Natalia Levshina*

Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages

DOI 10.1515/flin-2016-0019
Submitted May 25, 2015; Revision invited November 17, 2015;
Revision received February 9, 2016; Accepted May 31, 2016

Abstract: This paper investigates variation of lexical and analytic causatives in 15 European languages from the Germanic, Romance, and Slavic genera based on a multilingual parallel corpus of film subtitles. Using typological parameters of variation of causatives from the literature, this study tests which parameters are relevant for the choice between analytic and lexical causatives in the sample of languages. The main research question is whether the variation is constrained by one semantic dimension, namely, the conceptual integration of the causing and caused events, as suggested by previous research on iconicity in language, or whether several different semantic and syntactic factors are at play. To answer this question, I use an exploratory multivariate technique for categorical data (Multiple Correspondence Analysis with supplementary points) and conditional random forests, a nonparametric regression and classification method. The study demonstrates the importance of corpus data in testing typological hypotheses.

Keywords: parallel corpus, analytic causative, lexical causative, Multiple Correspondence Analysis, random forests

1 Introduction

Typology is a domain where quantitative corpus-based investigations are the exception rather than the rule. This does not mean that typological analyses never involve statistics; however, statistically informed analyses usually occur at the level of languages as observations in large typological databases or at the level of constructional types, e. g., inchoative–causative pairs in Haspelmath (1993) or argument microstructures in Hartmann et al. (2014). With a few notable exceptions, such as an exploratory analysis in lexical typology in Wälchli and

***Corresponding author: Natalia Levshina,** Leipzig University (IPF 141199), Nikolaistraße 6–10, 04109 Leipzig, Germany, E-mail: natalia.levshina@uni-leipzig.de

Cysouw (2012), data-driven induction of semantic roles (Cysouw 2014), a study of frequency effects in causal-noncausal verb alternations in Haspelmath et al. (2014), and a multivariate analysis of noun/verb ratios in Bickel et al. (2015), investigations at the token level are still quite rare. This can be explained by two main reasons. One is the traditional influence of structuralism, which has played a crucial role in the description of lesser studied languages and which does not consider usage frequencies relevant for the description of a linguistic system. The second reason is of a more practical nature: for a vast majority of existing languages, linguists only have recourse to very small samples of data, which lend themselves only to qualitative analyses.

This situation is deplorable because many, if not most, generalizations at the level of language are only an approximation. Many descriptive chapters in the *World Atlas of Language Structures* (Dryer and Haspelmath 2013), for example, identify tendencies rather than categorical choices. For instance, the chapter on the word order of nouns and adjectives describes French as a noun–adjective language and Polish as an adjective–noun language (Dryer 2013); alternative word orders are, however, also possible in each of these languages, depending on the adjectival class (French) or the function of the noun phrase (Polish).[1] Although such coarse generalizations are by all means useful for detecting broad regularities in the way languages are organized, token frequencies can reveal cross-linguistic tendencies and clines that cannot be detected at the level of types. For example, the quantitative token-based corpus analysis in van der Auwera et al. (2012) demonstrates that Dutch is intermediate between English and German with regard to impersonalization strategies, whereas their analysis at the level of types does not allow for such a conclusion. In a similar vein, the author's own study (Levshina 2016) shows that the functions of verbs of "letting" in Romance and Germanic languages are distributed in a way that suggests the presence of an areal North–South continuum. The differences between the neighboring languages are quantitative, rather than qualitative, and can be pinned down only at the level of tokens. Hopefully, with the recent development of massively parallel corpora (e. g., Cysouw and Wälchli 2007) and comparable corpora (e. g., Quasthoff et al. 2014), the situation will gradually change and quantitative corpus-based investigations will become more common in typological research.

This paper aims to demonstrate, at least partly, what typology and general linguistics can gain from the token-based approach. More specifically, this study is the first attempt, to the best of my knowledge, to operationalize and test well-

1 In Polish, the noun–adjective word order is preferable in set expressions and taxonomies, where the adjective limits the class of objects expressed by the noun. For example, Cognitive Linguistics is translated as *lingwistyka kognitywna*.

known typological hypotheses about the division of labor between different causative constructions. According to a popular classification that goes back to Comrie's work (e. g., 1981: Ch. 8), causatives can be subdivided into three large groups that differ in the degree of syntactic integration of the causing and caused events:

(1) analytic – morphological – lexical

In contrast to another article on causative constructions recently published in this journal (Levshina 2015), which assumed a semasiological perspective on analytic causatives, this paper takes an onomasiological perspective and focuses on the variation between analytic and lexical causatives. An illustration is the well-known example *cause to die* and *kill* from Fodor (1970), as in (2):

(2) a. *The sheriff caused Bill to die.*
b. *The sheriff killed Bill.*

In analytic causatives, such as *cause + to die* in (2a), the causing and caused events are represented as separate predicates. In lexical causatives, such as *kill* in (2b), these events are merged in one lexeme. In English, there are no productive morphological causatives, but they can be found in other languages: in Turkish, for example, causative verbs can be formed with the help of a suffix, e. g., *öl-* 'die' vs. *öl-dür-* 'kill'.

The semantic differences between analytic, morphological, and lexical causatives have traditionally been interpreted iconically: the more compact a construction, the more integrated the causing and caused events are (e. g., Comrie 1981: 164–167; Haiman 1983: 783–788). From this, it follows that analytic causatives represent less semantically integrated events than lexical causatives. Although the iconic approach has been influential, a more sophisticated account was presented by Dixon (2000), who lists eight semantic and syntactic parameters that are relevant for formal variation of causatives in typologically diverse languages.

Neither the iconic approach nor Dixon's multifactorial account has been tested quantitatively yet. The only type of evidence presented by linguists is individual examples, usually invented rather than taken from a corpus. However, it is clear that a reliable way of disentangling several parameters, as in Dixon's case, is possible only with the help of statistical methods and empirical data. Moreover, if these parameters are correlated, it would be impossible to tell from isolated examples whether the formal variation can be explained by one parameter or another.

The present study aims to fill this gap. More exactly, I want to find out whether the formal variation of causative constructions boils down to one dimension (i. e., conceptual integration of the causing and caused events), as predicted by iconicity theory, or whether there are multiple dimensions of variation, as suggested by Dixon's study. I will use corpus data from 15 Indo-European languages that represent the Germanic, Romance, and Slavic genera. In these languages, the semantic "labor" of expressing causation is divided mainly between lexical and analytical causatives, unlike in some other Indo-European languages (e. g., Armenian or Hindi), which have productive morphological causatives. Although there have been many studies of analytic and lexical causatives in European languages, to the best of my knowledge, there has been no comparison of the division of labor between them.

The data for this study come from a parallel corpus of film subtitles compiled by the author (Levshina 2017). In order to answer the research questions, I employ different multivariate statistical techniques. One of them is Multiple Correspondence Analysis (MCA) with supplementary points, which allows one to visualize the correlations between different semantic and syntactic parameters of variation and to investigate possible common dimensions of variation, as well as the way the constructional forms are mapped onto this common semantic space. To zoom in on the cross-linguistic differences in the way the languages "carve up" the semantic space, I use conditional random forests, a nonparametric method that is useful for predicting a response (here, analytic or lexical causatives) that depends on a high number of intercorrelated predictors. Conditional random forests estimate the relative importance of different variables for the use of analytic and lexical causatives in each language.

The remaining part of the paper is organized as follows. Section 2 presents the theoretical background, namely, the parameters of variation and the definition of analytic and lexical causatives as comparative concepts. Section 3 describes the data and variables. Section 4 presents the MCA, whereas Section 5 discusses the results of the random forest modeling in 15 languages. Finally, Sections 6 and 7 offer a discussion of the results and final remarks.

2 Theoretical background

2.1 Causative constructions: Parameters of variation

There have been several proposals with regard to the semantic features that are relevant for the division of labor between different causative constructions. For

example, Comrie (1981: 164–167), who formulated the abovementioned distinction between analytic, morphological, and lexical causatives, proposed that this formal continuum corresponds to the semantic continuum from less direct to more direct causation. More compact causative forms tend to express more direct causation. For instance, stabbing someone dead represents an instance of more direct causation than tampering with one's gun so that the owner dies in a duel.

From this one-factor explanation, one can move on to more complex theories. Haiman (1983), who used causative constructions from several languages to illustrate the Principle of Iconicity, discussed two types of situations, depending on the animacy of the Causee. With animate Causees, the semantic difference between more compact and less compact causatives is most commonly related to the degree of agentivity, volitionality, or control on the part of the Causee. Consider (3):

(3) a. *He made the children lie down.*
b. *He laid the children down.*

Following Haiman (1983: 784), (3a) is possible if the children are awake and respond to the Causer's command or request by performing the action themselves; they are agentive participants. In contrast, (3b) is appropriate when the Causee is non-agentive; for instance, when the children are asleep or unconscious or unwilling to comply so that the Causer is the main source of energy in bringing about the result. This account also agrees with Givón's, who sees the main difference between analytic and lexical causatives in the presence or absence of an agentive human Causee (Givón 2001: 75).

With inanimate Causees, exemplified in (4) after Haiman (1983: 784), the semantic contrast is somewhat different. In (4a), the Causer employs some unnatural force (e. g., magic or telekinesis) to produce the result rather than his or her own physical force, as in (4b).

(4) a. *I caused the cup to rise to my lips.*
b. *I raised the cup to my lips.*

Yet, regardless of the animacy of the Causee, the analytic constructions show that the Causer is less directly involved in the causation process, relying on some other force, i. e., the Causee's agency in (3a) or magic in (4a). The difference between the lexical and analytic causatives in each sentence pair lies in the degree of (in)directness of causation. Haiman's account is thus very similar to Comrie's.

The most elaborate is Dixon's (2000) approach to the formal variation of causatives. He introduces a scale of formal compactness, which includes four classes. The compactness decreases from lexical causatives to periphrastic causatives:

- Lexical causatives, e. g., *break*$_{TR}$ or *walk*$_{TR}$;
- Morphological causatives, e. g., tone change, reduplication, or affixation;
- Complex predicates, e. g., serial verbs, French *faire* 'make' + V_{INF}, or causative particles;
- Periphrastic causatives, where the causatives are represented by verbs that belong to separate clauses, e. g., French *laisser* 'let' + NP + V_{INF} or Portuguese *fazer* 'make' + (NP) + V_{INF}.

According to Dixon, the degree of compactness is correlated with the semantic and syntactic features which are shown in Table 1 (Dixon 2000: 76). If a language has two different causative forms, a more compact and a less compact one, they will differ along one or more parameters.

Table 1: Dixon's (2000) semantic and syntactic parameters of variation of causative constructions.

	More compact forms	**Less compact forms**
1.	Non-causal verb describing a state	Non-causal verb describing an action
2.	Intransitive (or intransitive and simple transitive) non-causal verb	Transitive (ditransitive) non-causal verb
3.	Causee lacking control	Causee having control
4.	Causee willing ("let")	Causee unwilling ("make")
5.	Causee partially affected	Causee fully affected
6.	Direct causation	Indirect causation
7.	Intentional causation	Accidental causation
8.	Causation occurring naturally	Causation occurring with effort

The ninth parameter Dixon discussed is involvement of the Causer in the caused event. Yet, Dixon did not find any correlations between this parameter and the degree of compactness. His analysis, however, is based on a limited number of examples per parameter and does not take into account possible correlations between the parameters, which might pose a problem. Imagine a language from which the following two causatives are taken, which are translated as follows:

(5) a. *The mother carried the sleeping girl to bed.*
b. *The mother had the girl go to bed.*

In (5a), the Causer (the mother) is acting directly, and the Causee does not have control over the caused state (i. e., the girl is in bed). In contrast, (5b) has an indirectly acting Causer and a controlling Causee who performs an action, i. e., going to bed. Therefore, at least three variables – directness of the Causer, control of the Causee, and the Aktionsart properties of the caused event – are perfectly correlated in this example. In addition, the Causer in (5a) is involved in the caused event herself, which is not the case in (5b). In order to disentangle all these parameters, one needs many diverse examples (corpus tokens) and quantitative techniques, which would allow one to estimate the impact of each parameter while controlling for all others.

Section 3 will show how the semantic and syntactic parameters can be operationalized as variables in a quantitative corpus-based study. The description will be based on Dixon's list because it includes the previously formulated parameters by Comrie, Givón, and Haiman. I will also add a number of additional variables that were found to be relevant for the variation of causative constructions in my previous research.

2.2 Analytic and lexical causatives as comparative concepts

The present paper employs the comparative concept approach to language comparison (Haspelmath 2010), which is not rooted in any particular linguistic theory. Comparative concepts do not have to be cognitively real(istic) and do not have to be part of the linguistic system of any language speaker. The only criterion is their usefulness for cross-linguistic comparison. This subsection offers a definition of analytic and lexical causatives as comparative concepts.

2.2.1 Analytic causatives

In this paper, I follow the definition of analytic causatives proposed in Levshina (2015). From a semantic perspective, an analytic causative designates a causative event, which involves a causing event and a caused event (state), as well as their participants, most importantly the Causer and the Causee. The Causer initiates or is responsible for the causing event, whereas the Causee is the entity that brings about the caused event (state). From a formal perspective, an analytic causative is a construction that consists of two verbs and their arguments. One verb represents in an abstract way the causing event, whereas the other verb represents the caused event. The order of the predicates is irrelevant, although it is usually iconic. The clauses should be closely

integrated: at least some arguments of the second verb should also be grammatically dependent on the first verb.[2]

This definition of analytic constructions is very broad and contains both monoclausal and biclausal analytic causatives, e. g., French *faire* followed immediately by V_{INF} and Russian *zastavljat* 'make' + (NP) + V_{INF}. In Dixon's (2000) classification (see Section 2.1), complex predicates roughly correspond to monoclausal causatives, and periphrastic causatives overlap with biclausal ones. There are several reasons for conflating the two groups. First, the mono-/biclausality distinction is problematic since linguists constantly observe degrees of syntactic fusion rather than a clear-cut dichotomy (Kulikov 2001: 887). Second, it seems unwise to include the highly fused French *faire* + V_{INF}, which serves as a textbook example of monoclausality, and exclude the closely related Portuguese construction *fazer* + (NP) + V_{INF} or Romanian *a face* + complementizer *să* + V_{SUBJ}, which are less syntactically fused and can be considered biclausal.

From a semantic perspective, the definition includes both factitive ("making" or "having") and permissive causation ("letting") (see Nedjalkov 1976: Ch. 3), as well as more specific types such as forceful causation (force someone to do something), curative causation (have something done by somebody), and authorization (permit someone to do something).

A representative list of constructions that meet these formal and semantic criteria in 18 European languages is provided in Levshina (2015). However, in such a comparative study, it is important not only to describe the constructions that meet the criteria but also mention those that do not. In spite of their undeniable relevance to the discussion of causative constructions in general, I excluded the following constructions from the analyses presented below:

- causative constructions where the caused state is expressed by a nominal phrase (e. g., *We need to force their cooperation*), prepositional phrase (e. g., *They'll have you on your knees*), or adjective (e. g., *He's driving me crazy*). Note that phrasal verbs, which consist of two elements, e. g., *switch the lights off* or *turn up the volume*, are treated here as lexical causatives (see Section 2.2.2);
- periphrastic constructions where the first and second clauses do not contain a shared argument (e. g., *He made it so that she stayed with him*);

2 VERB can be defined, following the well-known prototype approach by Croft (1991), as a representative of the word class that typically expresses predication (vs. reference and modification) and action (vs. objects and states). The concept WORD is notoriously difficult to define (see Haspelmath 2011), but, at least for the languages in question, it can be defined as an autonomous meaningful unit which represents formally "a segment string that cannot be interrupted by a free form without changing its meaning" (Haspelmath 2010: 666).

- periphrastic constructions where the first verb is semantically more specific than the verbs in the abovementioned constructions (e. g., *She tricked him into doing this*), i. e., it can be classified as an instrumental copula, in Nedjalkov and Sil'nickij's (1973) terminology;
- adhortative and other modal constructions (e. g., *Let us do it*!);
- constructions that express assistive causation (e. g., *I helped him repair his car*);
- idioms, where the original force-dynamic meaning is no longer actual (e. g., Fr. vulg. *va te faire foutre* 'fuck off');
- constructions that do not allow to identify the Causer and the Causee and where the structure has lost transparency as a result of grammaticalization or lexicalization. For illustration, consider (6), which is an example of the subjectless modal passive construction in Polish (von Waldenfels 2012: Section 3.6). The sentence is impersonal and no Causer can be identified.

(6) Polish, *Avatar*

Zobaczymy,	*czy*	*twoją*	*głupotę*	*da*	*się*	*wyleczyć*.
see.PFV.FUT.1PL	if	your	foolishness.ACC	give.PRS.3SG	REFL	cure.INF.PFV

'We'll see if your foolishness can be cured.'

2.2.2 Lexical causatives

Lexical causatives are those where the causing and caused events are merged in one lexeme. Examples are *kill* and *melt*$_{TR}$. Consider the example in (7):

(7) *He dropped the towel onto the floor.*

The sentence can be paraphrased as "He caused the towel to fall/let the towel fall onto the floor". The main semantic criterion is that the Causer brings about a change in location, state, or possession status of the Causee (i. e., the object of the construction) or that the Causer creates it (e. g., *write a book*). That is why, the verb *break*$_{TR}$ is considered a lexical causative here, but *hit* is not.

Intransitive verbs with self-causation (e. g., *He ran* = "He caused himself to run") will not be regarded as lexical causatives here. As for reflexive verbs, they will be treated as lexical causatives when the Causer affects him-/her-/itself, as in *He washes himself*. When there is no such causation, the verb is not treated as a lexical causative. Consider (8), where the subject is not affecting him-/herself and the meaning is passive. Moreover, in (8), the verb does not have a non-reflexive counterpart without the suffix *-s'*.

(8) Russian
Vy zabluždaetes'.
you be.mistaken.PRS.2PL
'You're mistaken.'

3 Data and variables

3.1 Corpus of film subtitles and data extraction

This study employs a parallel corpus of film subtitles compiled by the author (Levshina 2017) from different online repositories where users upload their translations of different films. This source of data has several advantages over other existing parallel corpora, such as translations of the Bible, fiction, legal documents, or the European Parliament proceedings. First, the language of film subtitles is meant to represent spoken discourse, especially spontaneous conversations. The language in subtitles is informal, contemporary, and not restricted to a particular professional domain. Second, subtitles are freely downloadable from online repositories and are available for many typologically diverse languages. Third, there is timing information, which helps one align the captions.

Some people might object to using this source of data. First, the translations may in principle run the risk of being influenced by the source languages. Note, however, that translationese is a potential problem for all other types of parallel corpora, which are widely used in contrastive linguistics and are gaining popularity in typology. Moreover, no significant linguistic difference has been observed between source and target English in film subtitles, which suggests that translationese is a less serious issue than one may think (Levshina 2017). Second, the strict formal spatiotemporal limitations of a caption appearing on the screen might influence the choice of linguistic structures by the translator in favor of more compact ones. However, a comparison of *n*-grams of English subtitles with *n*-grams of several registers from contemporary corpora of written and spoken English suggests that the language of subtitles is very similar to that of spontaneous informal conversations (Levshina 2017). Thus, subtitles represent a reliable source of typological evidence.

The films that were selected for this study were in various original languages and represented different genres. The list is as follows:

- *Avatar*: epic science fiction, USA, 2009
- *Das Leben der Anderen* (*The Lives of Others*): drama, Germany, 2006

- *El laberinto del fauno* (*Pan's Labyrinth*): dark fantasy. Mexico/Spain, 2006
- *La vita è bella* (*Life Is Beautiful*): tragicomedy/drama, Italy, 1997
- *Le fabuleux destin d'Amélie Poulain (Amélie*): romantic comedy, France, 2001
- *The Tourist*: romantic thriller, USA, 2010
- *Twilight*: vampire romance film, USA, 2008.

The choice of films was motivated by the availability of translations in all 15 selected languages. The files were in the SubRip (.srt) format and contained the captions and the times when each caption should appear on and disappear from the screen. These files were transformed into XML format and sentence aligned with the help of the software *subalign* (Tiedemann 2012).

From these files, I extracted analytic and lexical causatives according to the definitions of the corresponding comparative concepts given in Section 2.2. The procedure was as follows. First, all analytic causatives were detected. I began with a list of known constructions in each language and then identified their translations with the help of the alignment information. If there were new analytic causatives among the translations, they were added to the list of constructions. The procedure was repeated until no new constructions were found. Next, I retrieved all lexical causatives that occurred as translational equivalents of the analytic causatives. This particular procedure was chosen because lexical causatives are much more frequent than analytic causatives, and it would be practically impossible to analyze all lexical causatives occurring in the data. I am aware that the selected approach may create a sampling bias. However, as will be shown below, lexical causatives do not exhibit much semantic variation, so the sampling procedure is unlikely to distort the results.

As a result of this procedure, I obtained 325 causative situations, which were represented as rows in a table. The 15 languages were represented as columns. From the total number of the cells in the table, only 42.7 % contain either an analytic or lexical causative. This means that in the majority of cases, the translators resort to other means: these are extremely diverse, ranging from modal verbs to causal connectives and prepositions, and from resultative constructions of the type *make* X + Adj to omission of the causative meaning (cf. Nedjalkov and Sil'nickij 1973). Although these causative and non-causative expressions are by all means highly relevant to a general study of causative constructions, they are beyond the scope of the present paper.

The frequencies of analytic and lexical causatives in each language are shown in Table 2. On the basis of the data in Table 2, it can be observed that the frequencies of analytic causatives varied much more substantially across

Table 2: The frequencies of analytic and lexical causatives in the data set.

Genus	Language	Analytic	Lexical	Total
Germanic	Dutch	82	53	135
	English	96	72	168
	German	71	71	142
	Norwegian	87	67	171
	Swedish	68	62	130
Romance	French	116	57	173
	Italian	118	43	164
	Portuguese	75	71	179
	Romanian	71	79	150
	Spanish	77	63	140
Slavic	Bulgarian	43	86	129
	Czech	60	81	141
	Polish	25	72	97
	Russian	28	82	110
	Slovenian	28	77	105
Total		*1,045*	*1,036*	*2,081*

the various languages than the frequencies of lexical causatives. Italian and French have by far the highest frequencies of analytic causatives (118 and 116, respectively). The Slavic languages contain a much lower number of analytic causatives than the other genera, although Czech with 60 constructions is close to Germanic and Romance.

3.2 Semantic and syntactic variables

All causative situations were then coded for 13 semantic and syntactic variables. Nine of them are an operationalization of Dixon's typological parameters and four are additional variables. These semantic and syntactic variables are described below, and they are also listed in Table 3. Note that some adjustments had to be made to Dixon's (2000) original definitions in order to be able to describe the actual data adequately. One semantic parameter from Dixon's list could not be operationalized, namely, *Affectedness of Causee*, which distinguishes between fully affected Causees (e. g., destroy a house completely) and partially affected Causees (e. g., destroy a house partially, i. e., make some chips fall) (Dixon 2000: 67). Unfortunately, I could not find any information in the data that would allow me to make this distinction. This predictor is thus left out from the subsequent analyses.

Table 3: Parameters of variation of lexical and analytic causatives operationalized as variables.

	Variable	Abbreviation	Values	Expectations
1	Aktionsart of the caused event	*CdEvent*	"Nonaction"	Lexical
			"Action"	Analytic
2	Number of main participants	*NumPart*	"2"	Lexical
			"3"	Analytic
3	Control of Causee	*CeControl*	"Yes"	Analytic
			"No"	Lexical
4	Causee acting willingly	*CeVol*	"Yes"	Analytic
			"No"	Lexical
			"Undef"	No clear expectations
5	Making or letting	*MakeLet*	"Make"	Lexical
			"Let"	Analytic
6	Causer acting directly	*CrDirect*	"Yes"	Lexical
			"No"	Analytic
7	Causer acting intentionally	*CrIntent*	"Yes"	Lexical
			"No"	Analytic
8	Causer acting forcefully	*CrForce*	"Yes"	Analytic
			"No"	Lexical
9	Causer involved in caused event	*CrInvolved*	"Yes"	No clear expectations
			"No"	
10	Semantics of Causer	*CrSem*	"Anim"	Lexical
			"Inanim"	Analytic
11	Semantics of Causee	*CeSem*	"Anim"	Analytic
			"Inanim"	Lexical
12	Coreferentiality of Causer with other main participants	*Coref*	"Yes"	No clear expectations
			"No"	
13	Polarity	*Polarity*	"Pos"	No clear expectations
			"Neg"	

Because of the highly subjective nature of the semantic variables, I performed an inter-rater agreement experiment for a sample of causative situations with two other linguists. The agreement scores (Light's kappas) were 0.7 and higher, which allows me to conclude that the coding schema is reliable enough.

3.2.1 Variable 1: Caused events are actions or nonactions

This variable is an operationalization of the state/action parameter in Dixon (2000: 63), which describes the semantic properties of the non-causal verb that undergoes causativization. Since the causing and caused events are merged in lexical causatives like *kill* or *break*, it was impossible to apply

Dixon's original parameter directly. For the purposes of this case study, it was reformulated as a property of the caused event with two values: "Action" (activities, accomplishments, and most achievements in Vendler's 1957 well-known classification) and "Nonaction" (states, e. g., *be* and *believe*, and those achievements that represent a change of state, e. g., *die*). Consider examples from the corpus in (9):

(9) a. *At that moment, on a restaurant terrace nearby the wind magically* ***made*** *two glasses* ***dance*** *unseen on a tablecloth*. [two glasses dance = caused activity, coded as "Action"] (*Amélie*)
b. *They got steak? Bullshit*, ***let*** *me* ***see*** *that*. [I see that = caused state, coded as "Nonaction"] (*Avatar*)
c. *Don't shoot, you'll* ***piss*** *him* ***off***. [he becomes pissed off, i. e., angry = caused change of state, coded as "Nonaction"] (*Avatar*)

In line with Dixon (2000: 63), one can expect actions to boost the probability of analytic causatives and nonactions to increase the chances of lexical causatives.

3.2.2 Variable 2: Number of main participants

This variable is an adaptation of Dixon's (2000: 63–65) transitivity parameter. As with the previous variable, this adaptation is motivated by the absence of a non-causal verb in lexical causatives. All causative constructions were classified into constructions involving two main participants (e. g., *X causes Y to die*; *X kills Y*) and those involving three main participants (e. g., *X had Y kill Z*; *X gives Y Z*). For analytic causatives, the distinction between two-participant and three-participant causatives corresponds to the distinction between intransitive and transitive analytic causative constructions (Kemmer and Verhagen 1994). As for lexical causatives, the distinction between two-participant and three-participant causatives corresponds to the distinction between simple transitive and ditransitive verbs.[3] The theoretical basis for this operationalization is a proposal by Kemmer and Verhagen (1994) that

3 Ditransitives are defined as a comparative concept following Malchukov et al. (2010): they consist of a (ditransitive) verb, an Agent argument, a Recipient-like argument, and a Theme argument. Like Malchukov et al. (2010), I do not make the formal expression of the arguments more specific because there is substantial variation in the way that predicates and arguments are encoded in different languages.

transitive (three-participant) analytic causatives are semantically close to ditransitive verbs, whereas intransitive (two-participant) analytic causatives are close to simple transitive verbs. Following Dixon (2000), one can expect two-participant causatives to be more frequently lexical and three-participant causatives to be more frequently analytic.[4]

3.2.3 Variable 3: Control of the Causee

This variable corresponds to Dixon's (2000) distinction between Causees having control, as in (10a), and Causees lacking control, as in (10b):

(10) a. *Who gets **them*** [indigenous people] *to move?* (*Avatar*)
b. *I wanted to kill **him**, but I saw a sign from Eywa...* (*Avatar*)

One can expect analytic causatives to be used more frequently when the Causee has control, and lexical causatives when the Causee lacks control.

3.2.4 Variable 4: Causee acting willingly

An example with a willing animate Causee is given in (11a), whereas an unwilling animate Causee is exemplified by the context in (11b).

(11) a. *This is all because of that junk you let **her** read.* (*Pan's Labyrinth*)
b. *I wanted to kill **him**, but I saw a sign from Eywa. (Avatar)*

Coding this variable required a careful contextual analysis. In quite a few cases, no information was available in the context about the willingness of the Causee and the variable was coded as "Undefined". Although a Causee that has control (see Variable 3) is often a willing Causee, this is not always the case. Consider (12), where the Causee is in control of whether he will wait for the Causer or not but does not show willingness to wait:

(12) *Let's go, special case! Do not **make** me **wait** for you.* (*Avatar*)

---

4 It is important to mention here that implicit Causers (e. g., *Let me go*!) and implicit Causees (e. g., *I had my hair cut*) were always counted as main arguments. In contrast, demoted Affectees (Goldberg 2005), as in *His delusions made him kill*, were not counted as main participants.

3.2.5 Variable 5: "Making" or "letting"

Although Dixon (2000: 65–67) equates the parameter "willing vs. unwilling causee" with "letting" vs. "making", the present study will treat the latter distinction as a separate parameter. According to Talmy (2000: 419), letting is observed in cases when the Causee's intrinsic tendency is not overridden by the Causer. In the example *The plug coming loose let the water flow from the tank*, the Causee (water) has an intrinsic tendency to flow from the tank due to gravity; this tendency is not overridden by the Causer (the plug). The expectation is for letting to boost the probability of analytic causatives and for making to increase the probability of lexical causatives because "letting" designates less direct causation, where the Causer does not interfere in the natural cause of events. Note that Dixon (2000: 76) predicts the reverse: a more compact form for willing Causees and letting and less compact forms for unwilling Causees and making.

3.2.6 Variable 6: Causer acting directly

This variable distinguishes between direct and indirect causation. The following types of causation were considered indirect:

(i) Interpersonal causation by using communication, e. g., have someone make tea (= ask or order someone to make tea). According to Croft (1991: 166–167), this type of causation, which is labeled as inducive is normally indirect, since people cannot affect other people's minds directly, telepathy disregarded. An example is provided in (13a);

(ii) Causation involving natural or supernatural forces and technology, such as gravity (e. g., let something fall), magic (e. g., cause an object to float in the air by using telekinesis), or machines (e. g., fly an aircraft). An example is given in (13b);

(iii) Causation involving noninterference of the Causer (e. g., let someone be killed by not preventing it), as in (13c);

(iv) Causation where the effect follows the cause after a chain of intermediate events, as in (13d), so that there is no spatiotemporal overlap between the cause and the effect, as in (13d).

(13) a. German, *The Lives of Others*
Ich würde ihn ***überwachen lassen***.
I would him watch let
'I'd have him monitored.'
b. *He* ***dropped*** *me just before my act*. (*Amélie*)

c. *You must have **let** your phone **die** or something*. (*Twilight*)
d. *I can't bring myself to regret the decisions that **brought** me face to face with death*. (*Twilight*)

All remaining situations were considered cases of direct causation. Examples are given in (14):

(14) a. *The thrill of this rare contact **makes** her heart **beat** like a drum*. (*Amélie*)
b. *She likes **cracking** bones*. (*Amélie*)

In line with the previous studies, one can expect that less direct causation will increase the chances of analytic causatives in comparison with lexical causatives.

3.2.7 Variable 7: Causer acts intentionally

This variable distinguishes between intentional and accidental causation. Consider the examples in (15), where (15a) is an instance of intentional causation and (15b) exemplifies accidental causation.

(15) a. *I'll **roll** you into the mud*. (*Twilight*)
b. *And you are a moron that almost **ruined** my 8 million pound sterling operation*. (*The Tourist*)

Following Dixon (2000), one can expect intentional causation to increase the probability of lexical causatives, and accidental causation to be more associated with analytic causatives.

3.2.8 Variable 8: Causer acts forcefully

In some cases, causation may be fairly natural and involves no extra effort on the part of the Causer. An example is service encounters, such as in *He had his hair cut*. In other situations, the Causee may yield some resistance, and the Causer has to override this resistance by means of forceful causation, applying extra effort, as in *She was forced to retire early, although she still wanted to work*. In this case, the Causee is unwilling. Consider (16a), an example of forceful causation, and (16b), an example of non-forceful causation:

(16) a. French, *Avatar*
Je dois savoir comment les ***forcer à coopérer***.
I must know how them force to cooperate
'I must know how to force them to cooperate.'
b. *I* ***had*** *my nasal cavities* ***fixed***. (*Amélie*)

According to Dixon (2000), forceful causation is associated with less compact causative forms than "natural" causation.

3.2.9 Variable 9: Causer involved in the caused event

This parameter captures whether the Causer also performs the action specified in the caused event. For example, the Causer in *Bring your friends with you* is invited to come together with his or her friends. Dixon (2000) does not make any predictions on the effect of this parameter on the form of the causative. There were very few examples of involved Causers in the corpus; one such example is (17):

(17) Swedish, *Amélie*
Där mötte han några Afghanska kuppmakare... som ***fick***
then met he some Afghan coup-plotters who made
honom ***att stjäla*** *några ryska stridsspetsar*.
him to steal some Russian warheads
'There he met some Afghan conspirators ... who took him to steal some Russian warheads.'

3.2.10 Variable 10: Semantic class of the Causer

I also coded the Causer for the broad semantic class of animacy, where two values were identified, i. e., "animate" and "inanimate". Since intentionally acting Causers (Variable 7) are normally animate, one can expect animate Causers to increase the likelihood of lexical causatives.

3.2.11 Variable 11: Semantic class of the Causee

The Causees were coded for animacy as well, with two values "animate" and "inanimate". Since only animate Causees can have control over caused events, one can expect animate Causees to boost the chances of analytic causatives.

3.2.12 Variable 12: Coreferentiality of the Causer with the other main participants

The Causer may be coreferential with other main participants such as the Causee and the Affectee (i. e., the third argument of transitive analytic causatives and ditransitive lexical causatives, the end point of the causation chain). Examples of a coreferential Causer are *I let myself go* (Causer = Causee) and *I let them defeat me* (Causer = Affectee). There were no clear expectations with regard to the effect of this parameter on the form of the causative.

3.2.13 Variable 13: Polarity

This variable has two values: "positive" (*I made him do it*) and "negative" (*I didn't make him do it*). Negative polarity was coded when the clause with the causative construction contained a negative particle, pronoun, or adverb. There were no clear expectations with regard to this parameter. In the case of negation, the other semantic variables were coded on the basis of the non-negative equivalent of the original sentence. For example, *I didn't let him do it* was treated as *I let him do it* and thus was regarded as an instance of permissive ("letting"), rather than factitive ("making" or "having") causation.

3.3 Coding of the data set

The multilingual exemplars from the corpus were coded for the variables described in the previous subsection. The causative situations were first represented in terms of a formal description, which included the Causer, the Causee, the causing event, and the caused event. For example, the sentences in (18) were represented as follows: [the Post]$_{\text{CAUSER}}$ [CAUSES]$_{\text{CAUSING\_EVENT}}$ [a letter]$_{\text{CAUSEE}}$ [to arrive to DESTINATION]$_{\text{CAUSED\_EVENT}}$.

(18) *Amélie*

a. French

La Poste a le plaisir de vous ***faire parvenir***
the post has the Pleasure Of you make reach
la lettre ci-jointe.
the letter enclosed
'The Post has the pleasure of sending you the enclosed letter.'

b. *We are* ***forwarding*** *the enclosed letter to your address...*

In coding the utterances, the context of the utterance, including the visual and auditory information from the film, was taken into account.

In quite a few cases, different languages encoded the causative situations in different ways, so a general semantic representation was difficult to attain. For example, in (19), which has been taken from *Avatar*, the English version contains three participants (the Causer, the Causee, and the Affectee *this thing*, which is the direct object of *micromanage*, the predicate that specifies the caused event), whereas the Russian version has only two (the Causer and the Causee), since the effected predicate does not have a direct object.

(19) *Avatar*
  a. *I'm not about to* ***let*** *Selfridge and Quaritch* ***micromanage*** *this thing*.
  b. Russian
    Ja ne ***pozvolju*** *im* ***sovat'sja*** *v moj otdel*.
    I not allow.FUT them meddle in my department
    'I won't allow them to meddle in my department.'

In such situations, I assigned the most common value cross-linguistically, i. e., the one that was shared by the majority of the languages where this situation was expressed.

4 Common dimensions of variation of European causatives: Multiple Correspondence Analysis with supplementary points

This section discusses the results of a MCA with supplementary points. The method was introduced for cross-linguistic comparison in a study of English and Dutch analytic causatives by Levshina et al. (2013). In the present study, this approach is employed to carry out two tasks. First, it enables one to identify the most important semantic dimensions of variation shared by the causatives in the 15 European languages. These semantic dimensions, which in fact represent associations between the variables, form a common space of semantic variation of the causative constructions in the 15 languages. Second, it allows one to investigate the cross-linguistic differences in the way the linguistic forms can be mapped onto this space.

MCA is a technique that was designed for the analysis of multivariate categorical data. It can be viewed as an analog of Principal Component

Analysis for categorical variables. The data should be a matrix (table) with individual observations (in our case, causative situations) as rows and categorical variables (semantic and syntactic parameters) as columns. MCA is an exploratory dimensionality-reduction technique, which helps represent multivariate data in a small number of dimensions. These dimensions are expected to explain the variance, which is called inertia in MCA. It is measured on the basis of the Chi-square statistic, which reflects the difference between the real and expected co-occurrence frequencies of values of different variables. The first dimension of an MCA explains the most variance, the second, and subsequent ones less. The method also allows for representation of the data in low-dimensional maps. In this paper, I use the *ca* package (Nenadić and Greenacre 2007) in R (R Core Team 2015). More specifically, I employ adjusted MCA, which is a modification that gives a more realistic idea of how well the model fits the data than the "default" types of MCA based on the Burt and index co-occurrence matrices (Greenacre 2007: Ch. 19).

This section presents a two-dimensional MCA of the 13 semantic and syntactic variables, which can explain 67.1% of the total variance. The first dimension explains 56.1% and the second one 11%. The other dimensions contribute less than 5% and do not yield a clear semantic interpretation, so only the first two dimensions will be discussed. The map of the first and second dimensions is shown in Figure 1.

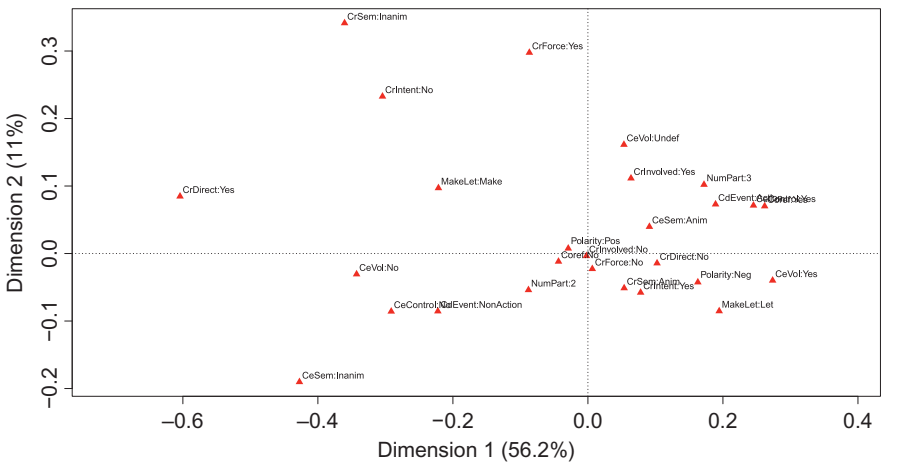

Figure 1: A Multiple Correspondence Analysis map based on 13 semantic and syntactic variables. Dimensions 1 and 2.

The map displays the specific values of the semantic and syntactic variables listed in Table 3, e. g., *CeControl:Yes* and *Coref:No*. The part of the label preceding the colon represents the categorical variable (i. e., *CeControl* and *Coref*), and the part following the colon represents the value (i. e., "Yes" and "No"). In the subsequent discussion, I will refer to these specific values of the variables as semantic and syntactic features. The location of the features with regard to one another can tell us how often they co-occur in the data.

The interpretation of the dimensions of variation, as represented by the axes in the plot, can be performed on the basis of the contributions of the semantic features to the orientation of the axes. The horizontal dimension is mostly determined by the feature *CrDirect:Yes*, which is located in the left part of the plot. It is followed by *CeVol:No* and *CeControl:No*, which are found on the left as well. The other contributions can be found in Table 4. This means that the horizontal dimension can be interpreted as (in)directness of causation, which is closely associated with volitionality and control of the Causee, as was shown in previous research (see Section 1). The part of the plot to the left of the vertical axis is thus associated with direct causation, and the right-hand area is associated with indirect causation.

As one can see from the contributions in Table 4, the vertical dimension is determined mostly by the animacy of the Causer, followed by the Causer's intentionality. This means that the area at the top of the map, above the horizontal axis, is associated with inanimate non-intentional Causers, and the area at the bottom with animate Causers, who usually (but not always) act intentionally.

Table 4: Contributions of semantic and syntactic features to Dimension 1 and 2.

Feature	Dim 1	Dim2	Feature	Dim 1	Dim 2
CrSem:Anim	6	29	*CeVol:No*	108	4
CrSem:Inanim	41	*189*	*CeVol:Undef*	1	61
CeSem:Anim	17	16	*CeVol:Yes*	80	*9*
CeSem:Inanim	78	79	*CdEvent:Action*	47	36
Coref:No	4	1	*CdEvent:NonAction*	55	42
Coref:Yes	24	9	*CrForce:No*	0	6
Polarity:Neg	10	3	*CrForce:Yes*	1	79
Polarity:Pos	2	1	*CrIntent:No*	46	136
NumPart:2	12	24	*CrIntent:Yes*	12	33
NumPart:3	24	43	*CrDirect:No*	22	2
MakeLet:Let	49	48	*CrDirect:Yes*	130	13
MakeLet:Make	56	55	*CrInvolved:No*	0	0
CeControl:No	95	42	*CrInvolved:Yes*	0	5
CeControl:Yes	79	34			

Most other variables are strongly associated with the horizontal dimension, which explains more than 50 % of the total inertia. On the left-hand side, which corresponds to direct causation, one can find inanimate and unwilling Causees (*CeSem:Inanim* and *CeVol:No*), Causees performing Nonactions (*CdEvent: NonAction*), and mostly short causation chains (*NumPart:2*). These features are located at the bottom. In the top-left quadrant, one finds directly acting Causers (*CrDirect:No*), factitive causation (*MakeLet:Make*), and forceful causation (*CrForce:Yes*). Moreover, the features that contribute the most to the vertical dimension, namely, unintentionally acting and inanimate Causers (*CrIntent:No*, *CrSem:Inanim*), are also located in the area of direct causation (at the top), which means that the two semantic distinctions are not exactly orthogonal.

The right-hand part of the map corresponds to indirect causation. The dispersion of the points is less wide than in the left-hand area, which suggests a stronger association between the features. The top-right quadrant contains the features that characterize a controlling and animate Causee (*CeControl:Yes* and *CeSem:Anim*, which partly overlaps in the plot with *Coref:Yes*). This Causee carries out an Action (*CdEvent:Action*) and affects another participant (*NumPart:3*), thus being an intermediate participant rather than the end of the causation chain. Also found here are coreferentiality (*Coref:Yes*), Causers involved in the caused event (*CrInvolved: Yes*), and Causees about whom one does not know if they act volitionally or not (*CeVol:Undef*).

In the bottom-right quadrant, very close to the horizontal axis, one can find indirectly acting Causers (*CrDirect:No*). Below are willing Causees (*CeVol:Yes*), negative polarity (*Polarity:Neg*), and permissive causation (*MakeLet:Let*), as well as the features associated with the vertical dimension, namely, intentionally acting and animate Causers (*CrIntent:Yes* and *CrSem:Anim*). Recall that volitionality of the Causee and making vs. letting are treated by Dixon as one parameter. One can see from the plot that these variables are indeed closely associated. However, they do not overlap completely.

Very close to the origin, one can find the following features: lack of coreferentiality (*Coref:No*), positive polarity (*Polarity:Pos*), lack of forceful causation (*CrForce:No*), and Causers who are not involved in the caused event (*CrInvolved: No*). All these features are by far the more frequent ones in comparison with their counterparts in the data set (namely, coreferentiality, negative polarity, forceful causation, and involved Causers). The features that are highly frequent tend to gravitate toward the origin in this Correspondence Analysis. A position far from the origin, in contrast, means that the feature is strongly correlated with the dimensions on the map.

After interpreting the first two dimensions of MCA, let us examine how these dimensions are associated with the causative constructions in the 15 languages.

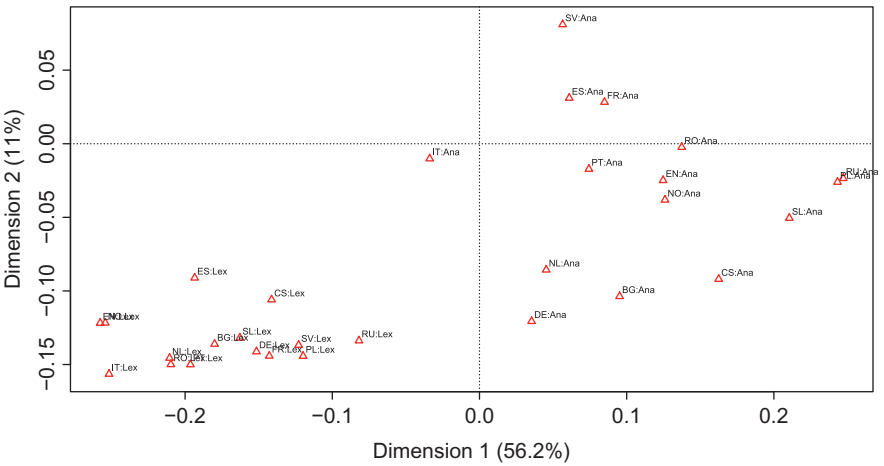

Figure 2: A Multiple Correspondence Analysis map based on 13 semantic and syntactic variables. Supplementary (passive) points. Dimensions 1 and 2.

Figure 2 shows the constructions, which are mapped onto the semantic space as supplementary, or passive points. Unlike active points (i. e., the semantic and syntactic features in Figure 1), supplementary points do not take part in the construction of the semantic space, being added only later to the space. Supplementary points are used when the variables are not homogeneous. It would be incorrect to add constructions (the forms) to the map as active points together with the semantic variables because they represent two different classes of phenomena (see Greenacre 2007: Ch. 18). To avoid overplotting, only the supplementary points are displayed. The positions of the language-specific constructions correspond to the average position of the cases (i. e., rows in the original data) that share the same construction.[5] For example, *IT:Ana* shows the average position of all observations with analytic causatives in Italian.

One can see that the lexical causatives form a dense cluster, which is located approximately in the bottom-left quadrant associated with direct and intentional causation (according to the horizontal and vertical dimensions, respectively). The analytic causatives are more widely distributed. This suggests that they are more diverse semantically than the lexical causatives. Still, the majority of the analytic causatives are located in the right part of the plot, which means that most of them tend to designate indirect causation with

5 Although the row labels are not shown here for the sake of visual clarity, the MCA algorithm in fact computes the coordinates for each row on the basis of the distribution of the semantic features.

autonomous Causees. However, the Italian analytic causatives are located quite close to the cluster of lexical causatives. This ties in with previous findings that Italian *fare* + V_{INF}, which is by far the most frequent analytic causative in the Italian subsample, is more grammaticalized than the corresponding constructions in the other Romance languages (Soares da Silva 2012; Levshina 2015). One can also observe that the Dutch and German analytic causatives (predominantly *laten* + V_{INF} and *lassen* + V_{INF}) are relatively close to the cluster of lexical causatives. Importantly, these three languages (Italian, Dutch, and German) have highly frequent and semantically broad analytic causatives that can represent both making and letting (Levshina 2015). This means that these causatives can function as simple valency-increasing devices without much semantic content. As such, they can be used to represent the semantics of tightly integrated events with nonautonomous Causees when a lexical causative is not available.

At this point, the following interim conclusions can be drawn. First, there are strong associations between variables, especially between those that relate to (in)directness of causation. However, not all variables are strongly associated with this semantic distinction; most importantly, these are animacy and intentionality of the Causer. This suggests that the semantic variation of causatives in the European languages is multidimensional. From the form–meaning mapping perspective, the cross-linguistic semantic variation of analytic causatives is greater than that of lexical causatives. Although most analytic causatives designate indirect causation, some analytic causatives (Italian, in particular, but also Dutch and German ones) are quite similar to lexical causatives. This may be a property of Standard Average European, since these languages belong to the core of the European linguistic area (cf. Levshina 2015).

In the next section, I will use random forests in order to reveal the similarities and differences between the languages with regard to the parameters that constrain the use of lexical and analytic causatives.

5 Zooming in on the individual languages: random forests

Random forests are a nonparametric regression and classification technique. They are popular in many scientific areas because they can cope with "small *n* large *p*" problems, highly correlated predictor variables and complex interactions. This technique is a perfect solution in our case. First, the data contain too many predictors in comparison with the number of observations (in particular,

the frequencies of analytic causatives in some languages are too low for a multiple logistic regression analysis). Second, many of these semantic variables are strongly associated, as was shown in Section 4, e. g., the animacy and control of the Causee, the number of participants and the semantic properties of the caused event. In linguistics, random forests have been successfully applied in variationist studies (Tagliamonte and Baayen 2012).

Random forests are "grown" from numerous individual classification and regression trees. In random forests, the algorithm draws several bootstrap samples from the original data set and creates a single classification tree for each sample. The trees in one forest will be different because of random variation in the samples. The prediction, which is then averaged across all trees, has been shown to be much more accurate than prediction based on single trees (Strobl et al. 2008). In this study, I use an approach based on conditional inference trees (Hothorn et al. 2006), which is superior to some other methods, such as the traditional CART algorithm, because the trees do not have to be pruned. This method is also unbiased with regard to the number of categories in a categorical variable.

In addition to measures of classification accuracy, one can use random forests to obtain variable importance scores, which reflect the role of each variable in predicting the outcome (in our case, the use of analytic and lexical causatives). It is a value that shows by how much the prediction deteriorates if one randomly reshuffles the values of a predictor. The stronger the association between the predictor and the response, the stronger the negative effect of such reshuffling on the predictive power of the model. This importance score is computed for each tree, and then averaged across all trees in a forest. Conditional variable importance is a special type of variable importance, which is similar to effects of variables in multiple linear regression: the effect of each variable is conditional on all other variables. This approach is implemented in the package *party* in R (Strobl et al. 2008).

I created a random forest from 1,000 trees for each language, and then computed the variable importance scores.[6] There were 14 predictors in each model: 13 semantic and syntactic variables and the films where the causatives occurred. This allowed me to filter out the possible individual biases of subtitle

6 The models reported in this section were fitted with the parameter mtry = 4, which defines the number of variables randomly sampled at each node in a classification tree. This corresponds to a rule of thumb, according to which the parameter value should be close to the square root of the total number of predictors. Note that mtry = 3 and mtry = 5 yielded very similar results, which are not reported here due to space limitations.

translators and effects of film genre and topic. The predictive power of the models was good, with the concordance index above 0.8 in all models.

The variable importance scores are presented in Figures 3–5 for each language. The dashed lines separate the relevant importance scores on the right from the irrelevant ones on the left. According to a rule of thumb, this cutoff value is the absolute minimum score (Tagliamonte and Baayen 2012: fn.14). One should keep in mind that variable importance scores cannot be compared across

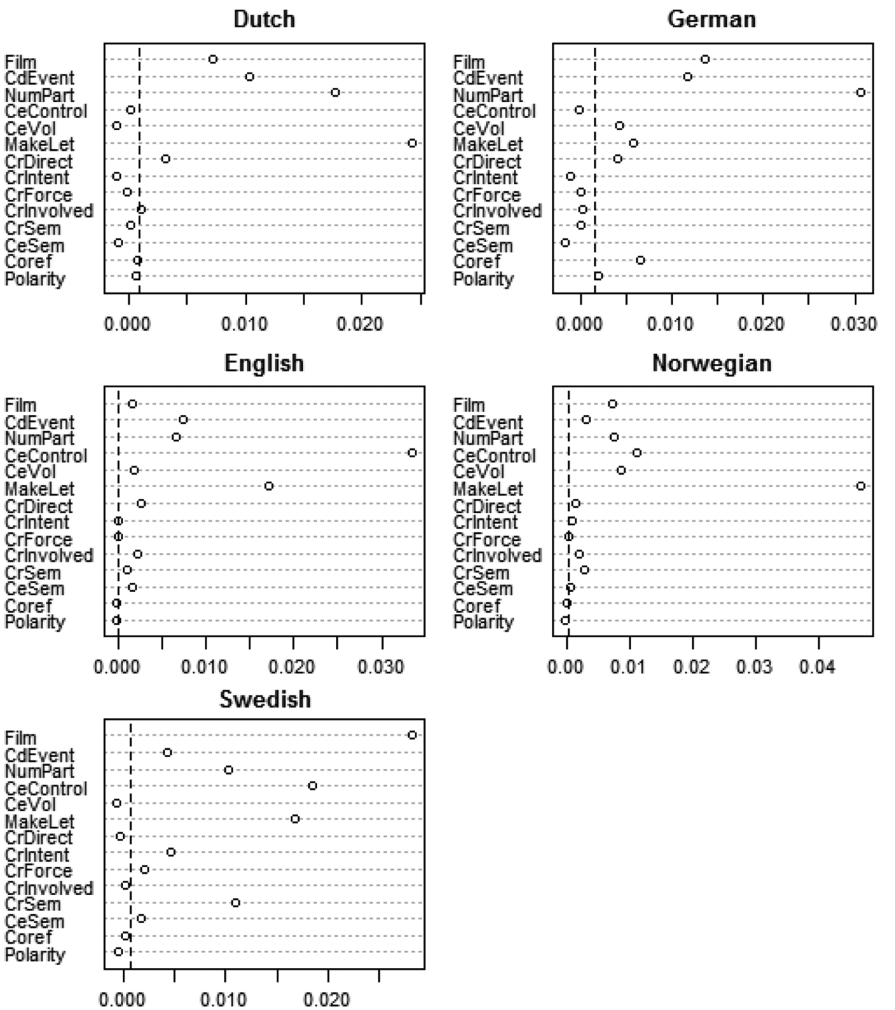

Figure 3: Conditional variable importance of semantic parameters in five Germanic languages, based on random forests.

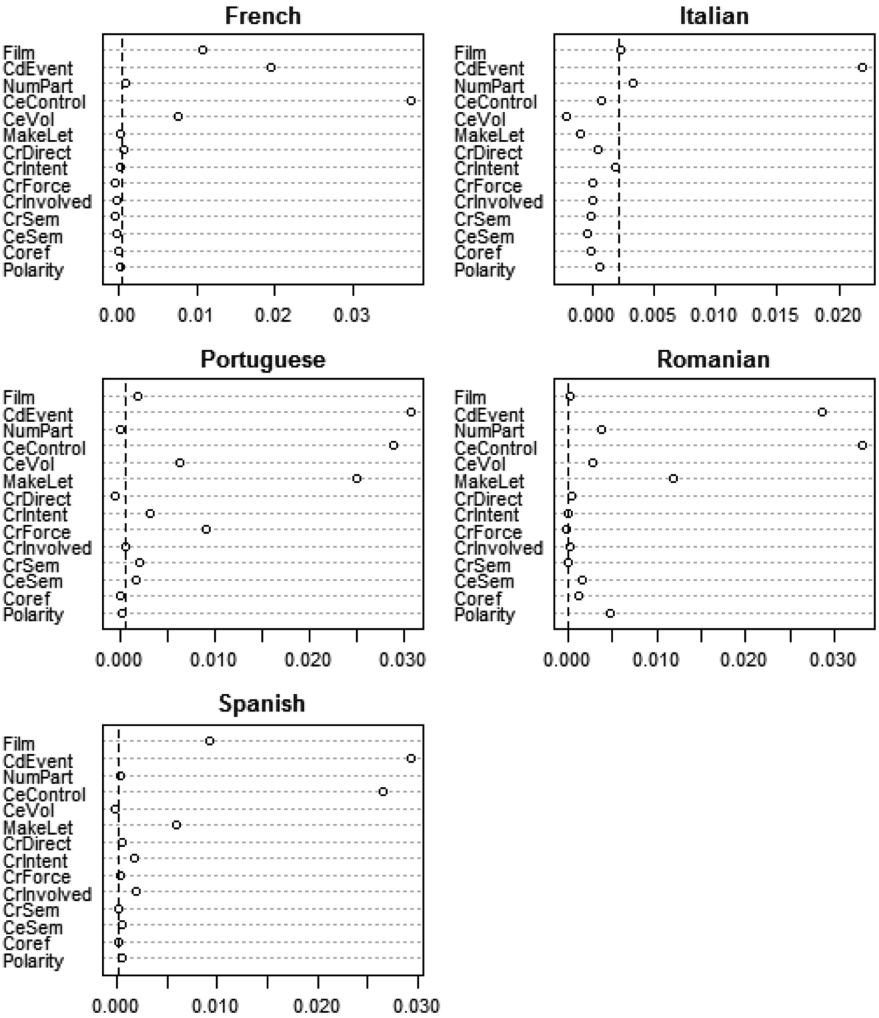

Figure 4: Conditional variable importance of semantic parameters in five Romance languages, based on random forests.

different languages. In what follows, I will examine which variables are the most prominent in each language and which are less important.

Figure 3 displays the conditional importance of the variables in the Germanic languages. The most prominent variables tend to be the distinction between making and letting, as well as the number of participants, directness of causation, control and volitionality of the Causee (as was shown in Section 4,

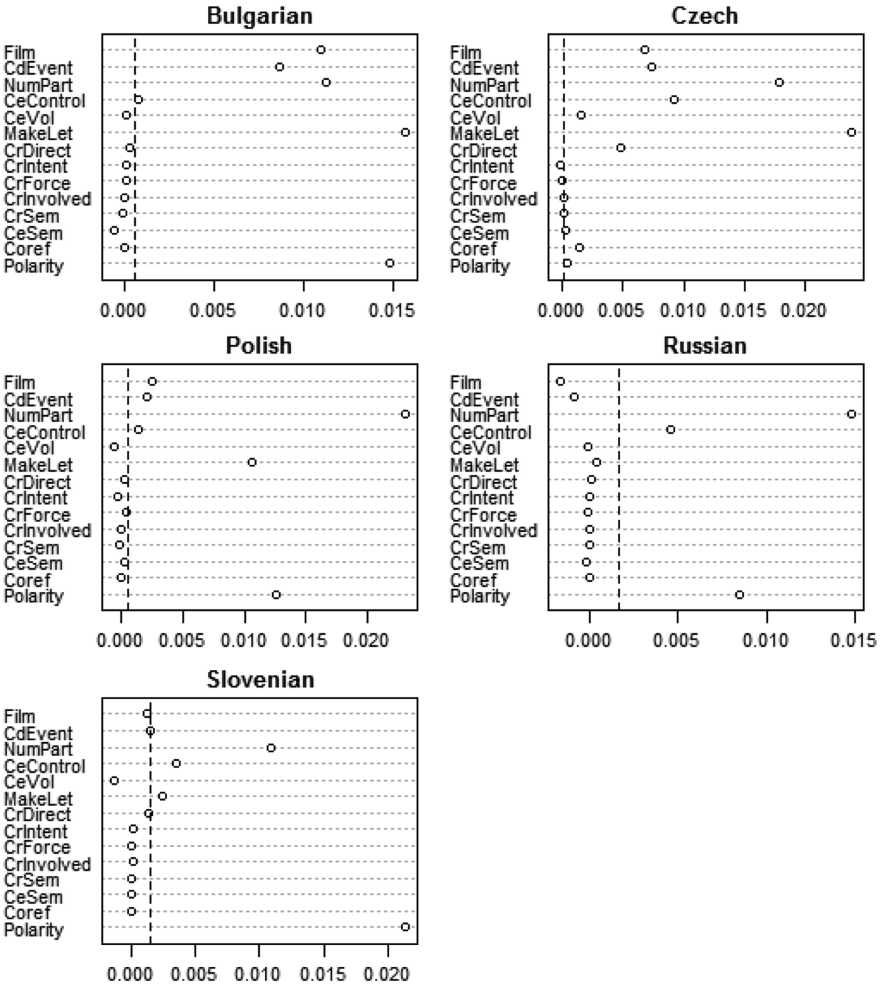

Figure 5: Conditional variable importance of semantic parameters in five Slavic languages, based on random forests.

these variables were mostly correlated with the horizontal dimension). However, the languages differ substantially with regard to the importance of individual variables. In English, the most prominent variable is control of the Causee (*CeControl*). In Swedish, this is the second most important variable after *Film*. The contrast between making and letting (*MakeLet*) is the strongest in Dutch and Norwegian. In German, the most important factor is the number of participants (*NoPart*). The features that are associated with the vertical dimension, namely,

intentionality and animacy of the Causer (*CrIntent* and *CrSem*) play a very modest role and matter somewhat only in Swedish. In addition, coreferentiality *Coref* is of some importance in German.

The Romance causatives display more homogeneous behavior, as Figure 4 suggests. In all five languages, the most important variables are the distinction between Actions and Nonactions (*CdEvent*) and/or control of the Causee (*CeControl*). These are followed in most languages by the distinction between making and letting (*MakeLet*). In Portuguese and Spanish, the intentionality of the Causer (*CrIntent*) plays some role. In addition, one should mention forcefulness of causation (*CrForce*) in Portuguese and *Polarity* in Romanian.

Finally, Figure 5 represents the Slavic languages. In all languages, some of the variables associated with the horizontal dimension play an important role, namely, control of the Causee (*CeControl*), making vs. letting (*MakeLet*), number of participants (*NumPart*), and the distinction between Actions and Nonactions (*CdEvent*). Again, there is a lot of variation. Making vs. letting (*MakeLet*) is the most important factor in Bulgarian and Czech. In Russian and Polish, the most important variable is the number of participants (*NumPart*). The variables associated with the Causer are unimportant in all languages. Coreferentiality (*Coref*) plays a visible role only in Czech, whereas *Polarity* seems to play some role in all languages. Its role is only marginal in Czech. In Slovenian, in contrast, coreferentiality is the most important variable. However, these results should be taken with a grain of salt because the frequencies of analytic causatives of most Slavic languages in the data set are very low. A larger scale study is needed to corroborate the results of these preliminary analyses.

The relative importance of *Film* varies greatly from language to language. This means that the effects of idiosyncratic preferences of constructions by the translators, as well as possible biases regarding the topic and genre, require a separate investigation.

To conclude, the semantic dimension of directness and indirectness of causation and the associated variable making vs. letting are the most important in all languages. The features associated with the dimension of intentionality and animacy of the Causer are less important (only to some extent in Swedish, Spanish, and Portuguese). Coreferentiality and/or polarity are only important in some Slavic languages, in German and Romanian.

6 Discussion

In all three genera, the semantic and syntactic parameters related to (in)directness of causation are the most prominent for the division of labor between

lexical and analytic causatives. This dimension and more specific semantic distinctions related to it (i. e., length of the causation chain, control and volitionality of the Causee, Actions vs. Nonactions as caused events, and making vs. letting) are related to the degree of integration of the causing and caused events. This means that the previous iconicity-related accounts by Comrie (1981), Givón (2001), and Haiman (1983), which were introduced in Section 2.1, hold generally. However, there are also other parameters that play a role and which cannot be easily interpreted in terms of these iconicity-related distinctions. Moreover, integration of the events can be expressed in very different ways cross-linguistically (as control of the Causee, number of the main participants, or making vs. letting, etc.). Even though these parameters are closely associated, it is still important to see which ones embody the underlying conceptual distinction the best. All this means that a multidimensional approach should be preferred because it gives a more exact picture.

It is interesting that Comrie mentions "true causation" (making) and letting as a possible semantic distinction but does not consider it important for explaining formal variation in the integration of the predicates (Comrie 1981: 164–165). However, the analyses show that this parameter is quite important grammatically, especially in some of the languages.

These findings should also be considered in light of Kulikov and Nedjalkov's (1992) hypothesis that causative markers that specialize in expressing permissive causation are unlikely. Indeed, analytic causatives as an abstract constructional type express both factitive and permissive causation in every language. However, the results of the analyses based on token frequencies suggest that analytic constructions in general are more often associated with letting and lexical causatives with making. Moreover, in every language in the sample, one can find specific analytic causative constructions (with such verbs as *let*, *allow*, and *permit* and their equivalents) that express only permissive causation.

Whether the observed form–meaning correspondence plays a direct causal role in determining the division of labor between lexical and analytic causatives is an important theoretical question. The fact is that analytic causatives are also normally longer than their lexical alternatives. Longer constructions also tend to be the less frequent ones, in accordance with the principle of economy (Haspelmath 2008). Therefore, there could be an alternative account of the isomorphism: the longer forms (analytic causatives) are associated with rarer functions, and the shorter forms (lexical causatives) with the more frequent functions. This is a hypothesis for future research. However, the results of the present study provide some indications in favor of the economy-related approach. The rich cross-linguistic variation in the semantic properties of analytic causatives (in contrast with the striking homogeneity of lexical causatives)

suggests that analytic causatives may indeed perform less frequent functions, and these peripheral functions may be very diverse cross-linguistically.

If one needs a multidimensional approach, is Dixon's (2000) approach useful? In general, the answer is positive, although one encounters some problems. First, it is not evident how to operationalize the predicate-related parameters for lexical causatives, since the latter do not contain a non-causal verb that has its own Aktionsart properties or transitivity. Second, Dixon's predictions concerning volitionality of the Causee and the distinction between making and letting do not seem to hold. Contrary to his expectations, unwilling Causees and making are associated with the shorter forms (lexical causatives), and willing Causees and letting with the longer forms (analytic causatives). This is not surprising, since letting expresses less direct causation than making with a weaker effect of the Causer, and willing Causees have a greater autonomy than unwilling ones.[7] Third, Dixon's list should be extended to include other parameters, especially the semantic class of the Causer, which codetermined the second dimension on the MCA map, as well as coreferentiality and polarity, which turned out to play a role in the variation of the causatives in some languages. Finally, one has to take into account the fact that many parameters are highly correlated. One needs therefore multivariate methods to establish the effect of each individual parameter, while controlling for the other parameters, as it was done in this study with the help of conditional random forests.

Obviously, the results are influenced by the way analytic and lexical causatives were defined as comparative concepts. This paper chose the most inclusive approach to analytic causatives. The usefulness of these definitions should be tested in the future on typologically diverse languages.

Another important question concerns the role of an available lexical causative for a given causative situation. One could expect it to influence the use of causative constructions. This question is not easy to answer because of rich synonymy in language. Consider, for instance, an analytic causative *cause* smth./smb. *to fall*. The historically related lexical causative *fell* smth./smb. would not be a perfect match because its semantics is restricted to causing trees to fall (Comrie 1981: 163). A better alternative might be *drop* smth./smb., but it contains some degree of unexpectedness and, moreover, has its own

7 This result is most likely due to the fact that inanimate Causees were coded as non-volitional. This may be against Dixon's original intentions, but his study gives no clear instructions with regard to inanimate Causees. One of the advantages of the token-based approach advocated here is that all possible cases have to be dealt with explicitly.

analytic causative *cause* smth./smb. *to drop*. To the best of my knowledge, this problem remains largely unexplored. Distributional semantic approaches based on large-scale corpus data could provide us with measures of semantic similarity between analytic and lexical causatives.

7 A final remark: Why should typologists care?

With a few exceptions, this study has by and large corroborated the theories which were based on individual examples. It shows the brilliance of individual linguistic intuition of the researchers. Does this mean that all the hard work involved in collecting and encoding the corpus data is superfluous? I believe the answer is negative, for the following reasons.

First, the fact that these theories are largely confirmed does not mean that this will be the case with any other theory. One can hope that this paper will inspire more quantitative tests of well-established and recent theories in typology and functional linguistics. Moreover, it has been shown here that not all variation boils down to one semantic dimension, and that the language-specific manifestations of this dimension can vary substantially.

Second, many, if not most, semantic and formal categories in language (as probably any other human categories) are not clear-cut, and the analyses at the level of types (usually based on native speakers' or language specialists' judgments), which involve categorical decisions, are in many cases not adequate. As Comrie mentions himself in his discussion of the semantic differences between lexical causatives, such as *kill*, and analytic causatives, such as *cause to die*, "it is difficult to invent situations where one or other of these expressions would be excluded, but it is easy to invent situations, and more especially pairs of situations, where one of the two variants is more appropriate than the other" (Comrie 1981: 166). In such situations, one clearly needs a probabilistic approach based on token frequencies.

Third, an investigation of different semantic and syntactic parameters which are highly correlated, like those in Dixon (2000), is only possible at the level of tokens in a multivariate quantitative study. It is impossible to interpret a formal contrast semantically based on only one or two examples when several semantic distinctions can be potentially involved.

Fourth, a quantitative analysis requires a very precise understanding of the semantic and syntactic parameters, thus exposing any vagueness or lack of objective criteria in the formulation of hypotheses. For example, this study has demonstrated that Dixon's predicate-related parameters (Aktionsart and transitivity) are not

directly applicable to lexical causatives. I was forced to reformulate these parameters in a way that can be tested on corpus data. One can hope that the growing popularity of quantitative approaches in typology will trigger a more rigorous formulation of linguistic hypotheses and will thus further linguistic theory.

Acknowledgments: The author is very grateful to Hubert Cuyckens, Karolina Krawczak, Michael Cysouw, and an anonymous reviewer for their invaluable criticisms and suggestions, as well as to Jose Garcia Miguel and Björn Wiemer for their consultations on some tricky language-specific constructions. I am also indebted to my ex-colleagues from the Catholic University of Louvain, Ludivine Cribble and Samantha Laporte, who helped me to obtain the measures of interrater agreement. The main part of this research was funded by a postdoctoral grant received from the Belgian research foundation F.R.S.-FNRS. All usual disclaimers apply.

Abbreviations

1/2/3 = first/second/third person; ACC = accusative; FUT = future; INF = infinitive; PFV = perfective; PL = plural; PRS = present; REFL = reflexive; SG = singular

Abbreviations used in figures

BG = Bulgarian; CS = Czech; DE = German; EN = English; ES = Spanish; FR = French; IT = Italian; NL = Dutch; NO = Norwegian; PL = Polish; PT = Portuguese; RO = Romanian; RU = Russian; SL = Slovenian; SV = Swedish

References

Bickel, Balthasar, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Taras Zakharko & Frank Seifart. 2015. Noun-to-verb ratio and word order. Presentation at the conference Diversity linguistics: Retrospect and Prospect, MPI EVA, Leipzig, Germany, 1–3 May.

Comrie, Bernard. 1981. *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.

Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: University of Chicago Press.

Cysouw, Michael. 2014. Inducing semantic roles. In Silvia Luraghi & Heiko Narrog (eds.), *Perspectives on semantic roles*, 23–68. Amsterdam: John Benjamins.

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie Und Universalienforschung (STUF)* 60(2). 95–99.
Dixon, R. M. W. 2000. A typology of causatives: Form, syntax and meaning. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Changing valency: Case studies in transitivity*, 30–83. Cambridge: Cambridge University Press.
Dryer, Matthew S. 2013. Order of adjective and noun. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/87. (accessed 20 March 2015).
Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info (accessed 20 March 2015).
Fodor, Jerry. 1970. Three reasons for not deriving "kill" from "cause to die." *Linguistic Inquiry* 1(4). 429–438.
Givón, Talmy. 2001. *Syntax: An introduction*, Vol. II. Amsterdam: John Benjamins.
Goldberg, Adele E. 2005. Argument realization: The role of constructions, lexical semantics and discourse factors. In Jan-Ola Östman & Miriam Fried (eds.), *Construction grammars: Cognitive grounding and theoretical extensions*, 17–43. Amsterdam: John Benjamins.
Greenacre, Michael. 2007. *Correspondence analysis in practice*, 2nd edn. Boca Raton, FL: Chapman and Hall/CRC Press.
Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.
Hartmann, Iren, Martin Haspelmath & Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language* 38(3). 463–484.
Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity*, 87–120. Amsterdam: John Benjamins.
Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.
Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3). 663–687.
Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.
Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog & Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625.
Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.
Kemmer, Suzanne & Arie Verhagen. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5(2). 115–156.
Kulikov, Leonid I. 2001. Causatives. In Martin Haspelmath, Ekkehard König, Wolfgang Oesterreicher, & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook*, 886–898. Berlin: Mouton de Gruyter.
Kulikov, Leonid I. & Vladimir P. Nedjalkov. 1992. Questionnaire zur Kausativierung. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 45(2). 137–149.

Levshina, Natalia. 2015. European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2). 487–520.
Levshina, Natalia. 2016. Verbs of letting in Germanic and Romance languages: A quantitative investigation based on a parallel corpus of film subtitles. *Languages in Contrast* 16(1). 84–117.
Levshina, Natalia. 2017. Online film subtitles as a corpus: An *n*-gram approach. To appear in *Corpora* 12(3).
Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2013. Mapping constructional spaces: A contrastive analysis of English and Dutch analytic causatives. *Linguistics* 51(4). 825–854.
Malchukov, Andrej, Martin Haspelmath & Bernard Comrie. 2010. Ditransitive constructions: A typological overview. In Andrej Malchukov, Martin Haspelmath & Bernard Comrie (eds.), *Studies in ditransitive constructions: A comparative handbook*, 1–64. Berlin: De Gruyter Mouton.
Nedjalkov, Vladimir. 1976. *Kausativkonstruktionen*. Tübingen: TBL.
Nedjalkov, Vladimir & Georgij Sil'nickij 1973. Typologie der kausativen Konstruktionen. *Folia Linguistica* 6(3/4). 273–290.
Nenadić, Oleg & Michael Greenacre. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software* 20(3). 1–13. Retrieved from http://www.jstatsoft.org/v20/i03/ (accessed 12 April 2015).
Quasthoff, Uwe, Dirk Goldhahn & Thomas Eckart. 2014. Building large resources for text mining: The Leipzig Corpora Collection. In Chries Biemann & Alexander Mehler (eds.), *Text mining – From ontology learning to automated text processing applications*, 3–24. Cham, Switzerland: Springer.
R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/ (accessed 12 January 2015).
Soares da Silva, Augusto. 2012. Stages of grammaticalization of causative verbs and constructions in Portuguese, Spanish, French and Italian. *Folia Linguistica* 46(2). 513–552.
Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9: 307. Retrieved from http://www.biomedcentral.com/1471-2105/9/307 (accessed 17 March 2015)
Talmy, Leonard. 2000. *Toward a cognitive semantics*, Vol. 1. Cambridge, MA: MIT Press.
Tagliamonte, Sali & R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 2214–2218. Istanbul. Available at www.lrec conf.org/proceedings/lrec2012/pdf/463 Paper.pdf.
van der Auwera, Johan, Volker Gast & Jaroen Vanderbiesen. 2012. Human impersonal pronouns in English, Dutch and German. *Leuvense Bijdragen* 98. 27–64
Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2). 143–160.
von Waldenfels, Ruprecht. 2012. *The grammaticalization of "give" + infinitive: A comparative study of Russian, Polish and Czech*. Berlin: De Gruyter Mouton.
Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.