Probability of external causation: an empirical account of cross-linguistic variation in lexical causatives

Abstract

This corpus-driven computational study addresses the question of why some verbs in some languages participate in the causative alternation while their counterparts in other languages do not. The results of this study suggest that the lexical property that underlies this variation is the probability of external causation. Alternating verbs are distributed on a scale of increasing probability for an external causer to occur.

The probability of external causation can be empirically assessed in two ways, among others: first, by observing the typological distribution of causative and anticausative morphological markings across a wide range of languages; second, through the frequency distribution of causative and anticausative uses of the alternating verbs in a corpus of a single language. Our study reveals that these two measures are correlated. Moreover, we demonstrate that the corpus-based measure is applicable to a wide range of verbs. Extending the corpus-based investigation comparatively across two languages, English and German, we find that the frequencies of cross-linguistic realizations of lexical causatives are modulated by the probability of external causation, an underlying parameter assigned to verb types. Finally, we propose a probabilistic graphical model that clusters verbs based on the relation between the cross-linguistic distribution of their causative and anticausative realizations and the probability of external causation.

Keywords: Causative alternation; Corpus distribution; Typological distribution; Bayesian modeling; Parallel corpora

1 Introduction

The causative (/inchoative) alternation has been recognized in the linguistic literature as a widespread linguistic phenomenon, attested in almost all languages (Schäfer 2009). This alternation involves verbs such as *break* in (1), which can be realized in a sentence both as transitive (1a) and intransitive (1b). Both realizations express the same event but specify the participants in the event differently. The transitive version specifies the causer of the event (*Adam* in [1a]), whereas the intransitive sentence does not. The transitive version is called causative, and the intransitive version is called anticausative. The verbs that participate in this alternation are commonly referred to as lexical causatives.

(1) a. Causative: *Adam broke the vase*.
b. Anticausative: *The vase broke*.

The causative alternation is an especially attractive topic of research because it allows insights into the underlying structure of the meaning of verbs. The ability to vary the number of expressed arguments while the verb stays the same is an indicator of a complex underlying structure of the verb that requires analysis and explanation. In addition, the causative alternation appears in different languages with a great variety of lexical, morphological, and syntactic realizations, which challenge the linguist's attempts at generalization. This study concentrates on the relation between the underlying lexical representation of the alternating verbs and the variation in the observable realizations of the alternation across languages.

First, variation is observed in the sets of alternating verbs across languages. Most alternating verbs are translation equivalents across many languages; however, some verbs only alternate in some languages while their translation equivalents in other languages do not. The verbs that do not alternate in some languages can be partitioned into two groups: those that appear only as anticausative (intransitive) and those that appear only as causative (transitive). Table 1 presents

examples of only anticausative and only causative verbs in English. As indicated by the examples, taken from the study by Alexiadou (2010: 178), the English verbs *arrive* and *appear* do not alternate: their transitive realization (causative in Table 1) is not available in English. Their counterparts in Japanese, or Salish languages, for example, are, however, found both as transitive and intransitive. Similarly, verbs such as *cut* and *kill* are only found as causatives in English, whereas their counterparts in Greek or Hindi, for example, can alternate between a causative and an anticausative use.[1]

Table 1: Availability of the alternation in some examples of verbs and languages (Alexiadou 2010:178).

	Causative	Anticausative
arrive, appear	+Japanese, +Salish, -English	+all languages
kill, cut	+all languages	+Greek, +Hindi, -English

Table 2: Morphological marking in some examples of verbs and languages (Haspelmath 1993: 91).

	Causative	Anticausative
Mongolian	xajl-uul-ax 'melt'	xajl-ax 'melt'
Russian	rasplavit 'melt'	rasplavit-sja 'melt'
Japanese	atum-eru 'gather'	atum-aru 'gather'

Languages also differ in the morphological realization of the alternation. Table 2 presents some examples of morphological variation, taken from the study by Haspelmath (1993: 91). In some languages, such as Russian, Mongolian, and Japanese, the alternation is morphologically marked. The morpheme that marks the alternation can be found on the intransitive form of the verb, while the corresponding transitive form is not marked (the case of Russian in Table 2). In other languages, such as Mongolian, the morpheme that marks the alternation is found in the transitive version, whereas the intransitive use is unmarked. There are also languages where both forms receive a causative or an anticausative marker, such as in the Japanese example shown in Table 2. English, on the other hand, is an example of a language where the alternation is not morphologically marked. The different marking strategies illustrated in Table 2 represent only the most common markings. Languages can use different options for different verbs. For example, anticausative versions of some verbs are not marked in Russian.

The variation in the availability of the alternation, as illustrated in Table 1, raises the question of why some verbs do not alternate in some languages. If the alternation is enabled by the underlying lexical structure, as suggested in the literature, all the verbs that describe the same events in different languages are expected to alternate. The variation in morphological marking (Table 2) is also puzzling: Why do languages not agree on which version of the alternating verbs needs to be marked? Is this variation free or are there some systematic patterns?

Finally, these two facets of variation also raise the question whether a connection exists between the availability of the alternation and morphological marking. Our study primarily addresses these questions.

[1] In the overwhelming majority of cases, the English verb *cut* can appear only as causative. Levin and Rappaport Hovav (2013: 54-56) discussed some cases of anticausative *cut*. They indicated that for *cut* to be anticausative, it must respect one of the main properties of anticausatives: the event must happen without the agent's continuous involvement. The examples proposed by them appear to be severely restricted lexically (they all have *rope* as subject), as they discuss in detail.

We examine the frequency distribution of the verb alternants in text corpora, showing that the variation in their text frequency, cross-linguistic availability, and morphological marking originates in an underlying parameter assigned to verb types, namely, the probability of external causation. Both within-language and cross-linguistic variation in the instances of alternating verbs are considered by analyzing data extracted from parallel corpora.

The lexical causative alternation, which we address in this study, is often discussed in the literature in relation to similar phenomena, which we do not address. On the one hand, lexical causatives participate in the syntactic causative alternation (*Adam broke the vase/Adam made the vase break*). This alternation has been extensively explored in the linguistic literature in the context of argument sharing in complex predicates (Baker 1988; Williams 1997; Alsina 1997; Collins 1997; Aboh 2009). On the other hand, the causative alternation is often compared with similar diathesis alternations such as the passive voice (*The vase was broken*) and the middle voice (*Porcelain vases break easily*), where the cause of an event can be left unexpressed (Reinhart and Siloni 2004; Marelj 2004; Alexiadou et al. 2006; Kallulli 2006, Alexiadou and Doron 2012, Alexiadou et al. 2015). In these approaches, generalizations about the causative alternation are formulated by comparing a large variety of constructions and their interpretative strength comes from their coverage.

Our study does not address these related phenomena but only the lexical causative alternation. We formulate generalizations without referring to other similar phenomena because our analysis relies on a large number of observations. It combines theoretical and typological accounts of the alternation with strongly expressive fine-grained probabilistic models of quantitative data gathered from large-scale language corpora.

2 Theoretical accounts of lexical causatives

A theory of the causative alternation mainly aims at discovering general formal mechanisms that underlie the alternation. Numerous theories have been proposed to explain the fundamental fact addressed in our study: many, but not all, verbs alternate. This fact raises the question of how the underlying structure of alternating verbs differs from the structure of non-alternating verbs.

These accounts share the assumption that events such as the one described by the verb *break* in (1a) comprise at least two subevents, where *Adam* is the argument of the first subevent (often labeled "CAUSE") and *vase* is the argument of the second subevent (often labeled "BECOME"). The analysis of sentences such as (1b) is less straightforward. Researchers have argued that the CAUSE subevent is present in the underlying representation of the event, but this is not explicitly expressed (e.g., Reinhart's (2002) analysis of lexical detransitivization); furthermore, they have argued that these realizations do not involve the CAUSE component (Horvath and Siloni 2011). We argue that the CAUSE component in the representation of the anticausative realizations of alternating verbs is a gradable feature that depends on the lexical content of the verb.

The most apparent common property of alternating verbs in different languages is that they describe an event in which the state of one of the participants (patient or theme) changes (Levin and Rappaport Hovav 1994): they are mostly change-of-state verbs. However, some alternating verbs do not describe a change-of-state event, such as *hang*, and some verbs describe a change of state but do not alternate. Verbs such as transitive *cut* and intransitive *bloom* do not alternate although their meaning involves a change of state, with the exception of a few specific cases.

These observations distinguish between external and internal causation, which defines more precisely what meaning component determines a verb's participation in the causative alternation. Verbs such as *hang* can alternate, although they do not describe a change-of-state event, because they describe a state that is externally caused (caused by an entity external to the hanging object [Levin and Rappaport Hovav 1994]). Verbs such as *bloom* do not alternate because the event of blooming is internally caused; the event of blooming does not require a causer external to the blooming entity. This generalization, however, does not cover all the cases. Additional constraints are proposed to account for verbs such as *cut* (externally caused but does not alternate, see footnote 1). These constraints include in their definition the notion of agentivity of the subject, that is, volitional involvement in the event, distinguishing between an agent and causer (Hale and Keyser

1993; Levin and Rappaport Hovav 1994; Reinhart 2002; Levin 2009) and the notion of specificity of the meaning of the verb that describes an event (Schäfer 2009; Haspelmath 1993).

Based on the cross-linguistic patterns of morphological marking and cross-linguistic availability of the alternation, Alexiadou et al. (2006) and Alexiadou (2010) proposed finer-gained distinctions and argued for four classes of verbs.[2] In addition to the classes of externally caused and internally caused verbs, a third group of "unspecified roots" and a fourth group of "agentive roots" are proposed. These classes are related to the distribution of morphological marking, where the verbs are divided into two types: those that mark the anticausative version (Class I in the proposal by Alexiadou [2010]) and those that do not (Class II). Morphological marking of anticausatives indicates external causation, whereas the lack of marking indicates either internal or unspecified causation. Cause-unspecified roots alternate in all languages, including languages such as English, while internally and externally caused verbs alternate only in some languages, that is, those equipped with a formal mechanism that enables such an alternation. Agentive verbs do not alternate. While this analysis relates the variation in the availability of the alternation and morphological marking in an innovative way, it does not prove to be robust in classifying new verbs not covered by the analysis. For instance, the Serbian verb *rasti*, which means "to grow," would be classified as cause-unspecified since its counterpart alternates in English. Since it alternates in English, it should alternate in Serbian too, but it does not. The lack of alternation in Serbian in this case cannot be explained by this account. Alexiadou (2014) further pursued the notion of internal causation. Alexiadou (2014) indicated that the class of internally caused verbs must be split into at least two subclasses associated with two distinct representations: typically internally caused verbs such as *blossom*, whose anticausative morphology in Greek is active, and verbs such as *ferment*, which share some properties with typically alternating verbs and whose anticausative morphology in Greek is non-active. However, Alexiadou et al. (2015) argued that the association between the lexical representation and morphology of verbs does not hold cross-linguistically. They compared three morphological types of anticausatives (marked, unmarked, and optionally marked) in Greek, Italian, French, and German and concluded that morphological marking on a verb is not "derivable" from the lexical properties of the verb.

An important precursor of the studies described below is the study by Haspelmath (1993). This study has been recently extended, and its results have been interpreted with a new explanation in the study by Haspelmath et al. (2014). Central to Haspelmath's (1993) approach is the notion of scale of external causation. A systematic typological analysis of the morphological marking in the two realizations of alternating verbs across a wide range of languages leads to the notion of scale: the transition between internal and external causation is gradual. In other words, each verb is both internally caused and externally caused to a certain degree. The degree of internal vs. external causation is reflected in morphological marking. The relation between the morphological marking and degree of causation cannot be observed in any single language (or a small set of languages), but it becomes apparent when a large sample of languages is observed.

In Haspelmath's (1993) view, there are several types of morphological marking of the alternation across languages.[3] Languages differ with respect to which type they prefer. Despite the different morphological preferences, a study of 21 pairs of alternating verbs indicates that certain alternating verbs tend to bear the same type of marking across languages. Verbs that are the lexical equivalents of English *freeze, dry*, or *melt* tend to be marked when used causatively in many different languages, whereas the equivalents of English *gather, open, break*, or *close* tend to be marked in their anticausative uses. Table 3 presents the distribution of morphological marking for all the verbs included in the study by Haspelmath (1993: 104). Note that the verbs are ranked according to the ratio between anticausative and causative marking. The verbs with a low ratio are found at the top of the table, and those with a high ratio are at the bottom.

[2] More precisely, Alexiadou (2010) referred to verb roots rather than verbs to emphasize the idea that the discussion concerns the lexical representation that is common to both the anticausative and causative realization of a verb. We keep the usual term "verb" to refer to the same entity.

[3] Certain authors (Piñón 2001; Alexiadou 2006a) argued against a direct derivation. For presenting Haspelmath's (1993) views, we keep the original morphological distinctions.

Table 3: Morphological marking across languages; A = anticausative, C = causative, E = equipollent – both versions marked, e.g., the Japanese verb in Table 2), L = labile (no formal difference between the two versions, such as in most English lexical causatives), S = suppletive (morphologically unrelated verbs function as alternants of one verb, e.g., *learn/teach*).

	Languages	A	C	E	L	S	A/C
boil	21	0.5	11.5	3	6	0	0.04
freeze	21	2	12	3	4	0	0.17
dry	20	3	10	4	3	0	0.30
wake up	21	3	9	6	2	1	0.33
go out/put	21	3	7.5	5.5	3	2	0.41
sink	21	4	9.5	5.5	1.5	0.5	0.42
learn/teach	21	3.5	7.5	6	2	3	0.47
melt	21	5	10.5	3	2.5	0	0.48
stop	21	5.5	9	3.5	3	0	0.61
turn	21	8	7.5	4	1.5	0	1.07
dissolve	21	10.5	7.5	2	1	0	1.40
burn	21	7	5	2	5	2	1.40
destroy	20	8.5	5.5	5	1	0	1.50
fill	21	8	5	5	3	0	1.60
finish	21	7.5	4.5	5	4	0	1.67
begin	19	5	3	3	8	0	1.67
spread	21	11	6	3	1	0	1.83
roll	21	8.5	4.5	5	3	0	1.89
develop	21	10	5	5	1	0	2.00
get lost/lose	21	11.5	4.5	4.5	0	0.5	2.56
rise-raise	21	12	4.5	3.5	0	1	2.67
improve	21	8.5	3	8	1.5	0	2.67
rock	21	12	4	3.5	1.5	0	3.00
connect	21	15	2.5	1.5	1	1	6.00
change	21	11	1.5	4.5	4	0	7.33
gather	21	15	2	3	1	0	7.50
open	21	13	1.5	4	2.5	0	8.67
break	21	12.5	1	2.5	2	0	12.50
close	21	15.5	1	2.5	2	0	15.50
split	20	11.5	0.5	5	3	0	23.00
die/kill	21	0	3	1	1	16	—

Haspelmath (1993:105) interpreted these findings as evidence in favor of a universal scale of increasing likelihood of spontaneous occurrence. The verbs with a low A/C ratio describe events that are likely to happen without the involvement of an agent or any external force. If the verb is used with an overt agent, the form of the verb contains a morphological marker in the majority of languages. The verbs with a high A/C ratio typically specify an agent; if the agent is not specified, the verb tends to get some type of morphological marking across languages. In this interpretation, the cross-linguistic A/C ratio is an observable and measurable indicator of a lexical property of verbs. It expresses the degree to which an agent or an external cause is involved in the event described by the verb. Conjecture 1 provides a summary of the notion of the scale of spontaneous occurrence.

Conjecture 1 (Haspelmath 1993)

The scale of spontaneous occurrence:

freeze > dry > melt >	.....	> gather > open > break > close
low A/C (spontaneous)		high A/C (non-spontaneous)

The notion of spontaneous occurrence can be related to the distinction between internally caused and externally caused events argued for in the other analyses. Both notions concern the same lexical property of verbs – the involvement of an agent or a causer in the event described by a verb. The events that are placed at the spontaneous extreme of the scale would be those that can be perceived as internally caused. The occurrence of an agent or an external cause in these events is unlikely. The probability of external causation increases as the verbs are placed closer to the non-spontaneous end of the scale. The notion of scale of spontaneous occurrence does not imply complete absence of the agent (or cause) in any event. It does not predict the ungrammaticality of uses such as **The summer's heat bloomed the flowers*. What follows from the notion of scale is that such uses are possible but unlikely.

In a recent study, like Samardžić and Merlo (2012) and Haspelmath et al. (2014), Heidinger (2015) revisited the notion of spontaneity. He proposed that the notion is "causalness." While the actual measures to quantify spontaneity and causalness are the same, the change of name indicates that this work aims to keep the investigation and interpretation at the level of usage patterns and interaction between form and function. Heidinger (2015) investigated whether the causalness of a verb affects the markedness of the encoding of causatives in French and Spanish. These languages have a four-way causative system, with marked and unmarked forms both in the causative and anticausative alternants of verbs that undergo the causative alternation. A balanced-sample corpus study of 20 French and 20 Spanish verbs indicates that both languages exhibit an interaction between causalness and encoding: high levels of causalness correlate with marked anticausative and unmarked causatives, whereas low levels of this property show the opposite pattern. An interesting finding of this work is that morphological encodings of causative meaning and syntactic causatives seem to behave in the same way with respect to frequency.

3 Our proposal

As more data from more languages are considered, the accounts of the distinctions between alternating and non-alternating verbs tend to posit finer distinctions and more classes of alternating verbs. While the new classes are well justified, it is unclear whether it is possible to classify all the verbs into a small set of distinct classes and how many classes should exist.

Our approach is an alternative to adding heterogeneous constraints to the proposed generalizations. We view external causation as a gradable property, which enables us to account for a whole range of different cases in terms of the extent of external causation. We follow Alexiadou (2010) in relating the pieces of evidence from a wide range of observations and Schäfer (2008) in including corpus data into the explanation. This study extends these approaches by considering large datasets using computational and quantitative methods. Our analysis results in a quantitative explanation for the cross-linguistic variation in the availability of the alternation. This explanation associates the cross-linguistic spread of the alternation – how many languages exhibit the alternation typologically – to the extent of external causation encoded as a (quantitative) lexical parameter – how likely the verb is to have an overt causer. This parameter applies at the level of lexical entry, the level of lexical representation that is common to both realizations of an alternating verb (verb type).

The methodology of our study draws on experimental studies that test the descriptive and predictive power of theoretical proposals. Computational models provide formal mechanisms to define the differences between classes of causative verbs in sentence processing (Stevenson and Merlo, 1997). These models demonstrate that formal properties of the classes of verbs that undergo the causative alternation generate detectable statistical indicators in a corpus and support automatic classification of the verbs in lexical semantic classes (Merlo and Stevenson 2001).

In our study, we draw on these crucial findings, which demonstrate that the lexical properties of verbs are reflected in the way they are used in a corpus. As in these studies, we consider frequencies of certain uses of verbs as an observable and measurable property, which serves as evidence of the underlying lexical properties. We explore this relation further, relating it to a deeper level of theoretical semantic analysis of verbs and to the cross-linguistic variation observed in typological surveys of grammatical features (at the level of type) and in parallel corpora (at the level of token).

We consider Haspelmath's (1993) scale as a basis for our experiments, addressing the questions raised in the introduction and developed in the literature review: how can we characterize the variability in the causative alternation across verbs and across languages? How can we characterize the variability in the causative morphological marking across verbs and across languages? Is there a connection between the two? We empirically test the proposed scale against corpus data, extend its empirical validation to a larger set of verbs, demonstrate its effects across languages in the pragmatically and semantically controlled environment of translation, and propose a model to automatically partition verbs into the lexical-semantic classes proposed in the literature.

4 Experiments

Our approach to lexical causatives is based on the statistical analysis and the modeling of large datasets. Assuming that the use of verbs is related to their lexical and grammatical properties, we investigate the distribution of the causative and anticausative realizations of a large number of verbs collected from a corpus of around a million and a half syntactically analyzed sentences. We identify the underlying lexical semantic properties of verbs that generate this distribution.

In the first experiment, we demonstrate that the distribution of the realizations of the alternating verbs in a corpus is correlated with the distribution of morphological marking across languages. We measure the correlation for a sample of 29 verbs for which typological descriptions are available (Haspelmath 1993). The work here extends the results reported in the study by Samardžić and Merlo (2012), where this correlation was reported for the first time in a published form.[4]

Regarding this correlation as a piece of evidence that the two distributions are generated by the same underlying property of the alternating verbs, we define this property as the degree of involvement of an external causer in an event described by a verb. We call this property *Sp* for spontaneity (following Haspelmath [1993]), but our measure does not directly correspond to the intuitive meaning of the word, as discussed in more detail below. We see the degree of *Sp* as a general scalar component in the lexical representation of the alternating verbs, assigned as a property to the lexical unit (the verb type, as opposed to the verb token) that is common to both realizations.

[4] This work predates both Haspelmath et al. (2014) and Heidinger (2015) in open-access, peer-reviewed, published form and we consider it to be the first instance of this line of work.

The value of this component, represented numerically, determines the observable behavior of all the verbs that participate in the alternation in any language.

Showing that the corpus-based measure of *Sp* is correlated with the typological measure enables us to extend the account to a larger sample of verbs. Since the corpus-based value is assigned to the verbs automatically, it can be readily calculated for practically any given set of verbs, replacing the typology-based value for which the data are difficult to collect. In the second experiment, we calculate the corpus-based *Sp*-value for 354 alternating verbs in English (Levin 1993). Using a statistical test, we show that the smaller set of verbs (the 29 verbs for which we measured the correlation) is a proper sample of the bigger set (the 354 verbs from Levin [1993]). This implies that the correlation established for the smaller set applies to the bigger set as well.

The fact that the verbs that are translations of each other alternate in some, but not all, languages requires explanation, as indicated in the literature. We assume that the *Sp*-value is involved in this differential availability of the causative alternation across languages. To examine the influence of *Sp*-value on cross-linguistic variation, in the third experiment, we analyze the distribution of causative and anticausative realizations in German translations of English lexical causatives. Through a statistical analysis of parallel instances of verbs, we identify certain trends in the cross-linguistic variation that are due to the *Sp*-value. In particular, we find that an overall tendency toward parallel translations (causatives are translated into causatives and anticausatives into anticausatives) is modulated by the *Sp*-value of the verb type. If the *Sp*-value was not the underlying source of corpus-usage behavior, we would find that all groups of verbs show the same tendency, irrespective of their *Sp*-value, with a preference for parallel translations. Instead, we find a tendency to regularize toward a construction congruent with the *Sp*-value (low *Sp* is congruent with anticausative and high *Sp* with causative uses). If a realization in one language is not congruent with the *Sp*-value, its corresponding realization is changed to agree with the *Sp*-value, thereby causing cross-linguistic variation. Mid *Sp*-value verbs do not show any regularization effect. They show a tendency toward parallel translation and are, therefore, not affected by cross-linguistic variation.

Based on these findings, we apply a modeling and classification methodology in the fourth experiment. We design a probabilistic model that exploits the information about the cross-linguistic variation to assess the *Sp*-value of lexical causatives, abstracting away from potential language-specific biases. The *Sp*-value is modeled as a latent variable in a probabilistic model of the bilingual causative alternation data, an unobserved causal entity that mediates the relation between the verb class and corpus usage. It is shown that this latent variable model can discriminate the causative and anticausative classes with good accuracy.

4.1 Experiment 1: Corpus-based validation of the scale of spontaneous occurrence

Haspelmath (1993) did not discuss the potential relation between the likelihood of spontaneous occurrence of an event and frequency of the different uses of the verb that describes this event in a single language. Nevertheless, it is logical to suppose that such a relation exists since the indicator of the likelihood, the morphological marking on the verbs, is considered to be a consequence of the way the verbs are used in general.

The placement of an event described by a verb on the scale can be expected to correspond to the probability for the verb to be used as a causative, hence transitively, or as an anticausative, hence intransitively, in any single language. On the other hand, the ratio of the frequencies of intransitive to transitive uses of verbs in a single language can be influenced by other factors as well, which can result in cross-linguistic variation. Note that the causative alternation is realized in different ways across languages: some languages mark the causative use of a verb, some mark the anticausative use, some mark both, and some mark neither of them (see tables 2 and 3). Morphological markers themselves can be special causative morphemes, but they can often be morphemes that also have other functions, such as the reflexive anticausative marker in most European languages. These factors might influence the ratio of intransitive to transitive uses in a given language.

To empirically validate the hypothesis that alternating verbs can be ordered on the scale of spontaneous occurrence of the events that they describe, we test it on corpus data. More precisely, we test the hypothesis that the distribution of morphological marking on the verbs across languages

and the ratio of their transitive to intransitive uses in a corpus are correlated. We can expect this correlation on the basis of the well-established correspondence between markedness and frequency (discussed in more detail by Haspelmath [2008]). In general, marked forms are expected to be less frequent than unmarked forms. Therefore, we expect the verbs that tend to have anticausative marking across languages to be used more often as causative (transitive) and those that tend to have causative marking to be used more often as anticausative (intransitive).

We calculate the ratio between the frequencies of causative (active transitive) and anticausative (intransitive) uses of verbs in a corpus of English for the verbs for which Haspelmath's study provides the typological A/C ratio, as shown in (2).[5]

(2)

$$C/A(verb) = \frac{frequency\ of\ causative\ uses}{frequency\ of\ anticausative\ uses}$$

We then measure the strength of the correlation between the ranks obtained by the two measures.

4.1.1 Materials and methods

Several clarifications are needed regarding the differences between the verbs used in Haspelmath's (1993) study and those in our study. This includes noting the criteria used to exclude some verbs. Most verbs analyzed by Haspelmath are also listed as participating in the causative alternation by Levin (1993) (e.g., *freeze, dry, melt, open, break, close*). Some verbs are not listed by Levin (e.g., *boil, gather)*. Nevertheless, we include them in the calculation because they clearly alternate. Four entries in Haspelmath's list are not English alternating verbs but converse pairs: *learn/teach*, *rise/raise*, *go out/put out*, and *get lost/lose*. We treat the former two pairs as single verb entries, adding the counts of the occurrences of both members of the pair. We do not calculate the ratio for the latter two because automatic extraction of their instances from the corpus could not be done using the methods already developed to extract the other verb instances. We exclude the verb *destroy* because it does not alternate in English and no complement verb is proposed by Haspelmath. Finally, the pair *kill/die* is excluded because its typology-based ranking is not available. This leaves us with 27 verbs for which we calculate the corpus-based C/A ratio.

Transitive, intransitive, and passive instances of the verbs were extracted from the English side of the parallel corpus Europarl (Koehn 2005), version 3, which contains around 1,500,000 sentences for each language. Syntactic relations needed for determining whether a verb is realized as transitive (with a direct object) or as intransitive (without object) are identified on the basis of automatic parsing using the MaltParser, a data-driven system for building parsers for different languages (Nivre et al. 2007).

Each instance is represented by a triple: the verb, head of its subject, and head of its object (if there is one). An English causative use of a verb is identified as an alternating verb realized in an active transitive clause. The anticausative use is identified as an intransitive use of an alternating verb. Passive is identified as the verb used in the passive participle form and headed by the corresponding passive auxiliary verb. Identification of the form of the clause that contains a lexical causative is performed automatically using the algorithm shown in Algorithm 1.

We are aware that it is a simplification to consider all the transitive uses of the alternating verbs as causatives and the intransitive uses as anticausatives. It can happen that a verb alternates in one of its senses but not in another. For instance, the sentence in (4) is not the causative counterpart of (3) but only of (5).

[5] Notice that we calculate the C/A ratio and not the A/C ratio, as Haspelmath (1993) does. Due to the inverse relation between marking and frequency, calculating the corpus score in the same direction as Haspelmath's measure would result in a negative correlation. We opt for a positive correlation to make the discussion easier to follow.

(3) *Mary was running in the park this morning.*

(4) *Mary was running the program again.*

(5) *The program was running again.*

Through a brief manual inspection of the lexical entries of the verbs in the Proposition Bank (Palmer et al. 2005), we assessed that this phenomenon is not frequent and that it should not influence the results. In our sample, only the verb *freeze* proved to be influenced by this phenomenon. This verb was discarded as an outlier while calculating the correlation between corpus-based and typology-based rankings of the verbs, but this was the only such example in the sample of verbs.[6]

Algorithm 1: Identifying transitive, intransitive, and passive uses of lexical causatives.

Input : 1. A corpus *S* comprising sentences *s* parsed with a dependency parser
2. A list of lexical causatives V

Output : The number of transitive, intransitive, and passive instances of each verb $v \in V$ in the corpus *S*

```
for i = 1 to i = S do
    for j = 1 to j = V do
        if v_j in s_i then
            if there is SUBJ that depends on v_j then
                if there is OBJ that depends on v_j then
                    return transitive;
                else
                    if v_j is passive then
                        return passive;
                    else
                        return intransitive.
```

As can be seen in Algorithm 1, only those instances where all the arguments are realized in the same clause were considered. This result is obtained by collecting only verb instances that have a subject. We exclude those realizations of verbs where either the subject or the object are moved or elided to control for influences of the specific syntactic structures on the interpretation of the meaning of verbs. Single clause realizations can be considered to be the typical and the simplest case. Although they are basically transitive realizations, the passive instances are separately extracted because the difference between active and passive transitive uses is crucial for the causative alternation, as discussed in detail by Alexiadou et al. (2006). Expressing the external causer – through a prepositional complement – is optional in passive constructions, while the external causer is obligatorily expressed as the subject in active transitive instances.

4.1.2 Results and discussion

We assess the strength of the rank correlation between the rank of the verbs based on the corpus-based C/A ratio and that based on the A/C ratio calculated on the typology of the morphological marking of the verbs across languages.

[6] The verb *freeze* is frequently used in our corpus in its non-literal sense (e.g., *freeze pensions, freeze assets*), while the sense that was considered by Haspelmath (1993) is most likely the literal meaning of the verb (as in *the lake froze*). Thus, the verb's corpus-based ranking was different from its typology-based ranking.

We obtain the Spearman rank correlation score $r_s = 0.67$, $p < 0.01$, with one outlier removed (the verb *freeze*, as discussed above). The score suggests a good correlation between the two sources of data. Figure 1 illustrates the scattergram representing the correlation. The coefficient of the correlation is strong enough to be taken as an empirical confirmation of Haspelmath's (1993) hypothesis. Given that the two distributions are significantly correlated, it is reasonable to assume that the same factor that underlies the typological distribution of morphological marking on verbs underlies the distribution of their transitive and intransitive realizations in a monolingual corpus. The property on which the two distributions depend is the probability of occurrence of an external causer in an event described by a verb. This property, with different (numerical) values, is assigned to each verb type, as part of the verb's lexical representation.

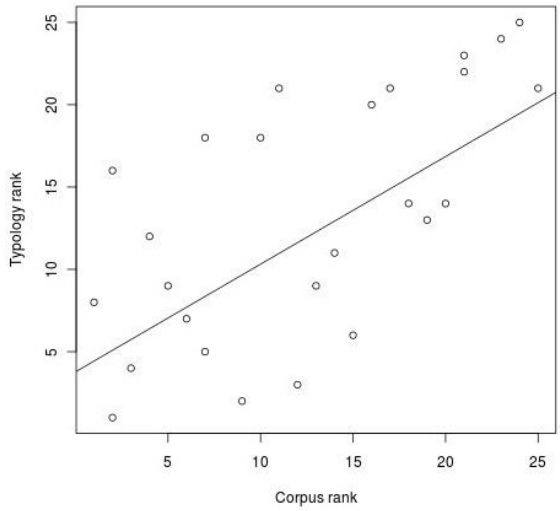

Figure 1: The correlation between the rankings of verbs on the scale of spontaneous occurrence.

Haspelmath et al. (2014) established a similar relation for six more languages, independently of Samardžić and Merlo (2012), confirming our result.[7] They provided, however, a different interpretation. While we propose that the above correlation is the effect of an underlying (and not directly observable) common property realized by the two measures, Haspelmath et al.'s (2014) interpretation of the result was that the relation between form and frequency is a direct, unmediated correlation between two observed variables, developed by pressure for efficiency. A system where the most frequent forms are the shortest, called a Huffman code in communication theory, is provably the most efficient code, the code that requires the smallest number of bits to encode a message. Haspelmath et al. (2014) proposed an explanation in terms of efficient coding: the less frequent forms are encoded as long forms (they have morphological marking), whereas the most frequent forms are encoded in short forms. The difference between the two interpretations is discussed in more detail in the general discussion, after presenting the results of our remaining three experiments, which provide more evidence in support of our interpretation.

4.2 Experiment 2: Scaling up

The correspondence between the automatically obtained corpus-based ranking of verbs and the scale of spontaneous occurrence confirms Haspelmath's theoretical hypothesis. This also means that

[7] In fact, our results are not replicated for English. Haspelmath et al. (2014) did not offer an explanation for the different results. We believe that the difference might be caused by the fact that the analysis by Haspelmath et al. (2014) is performed manually on a small sample counting passive instances as causatives, while we analyze a large sample automatically and exclude passive instances from the calculation.

the *Sp* feature can be calculated automatically from corpus data. To test whether the correlation that we find for the small sample of verbs discussed in Section 4.1 applies to a larger set, we compare the distribution of the corpus-based measure of *Sp* over this sample and the distribution of the same value for the 354 verbs listed by Levin (1993).

4.2.1 Materials and methods

The list of English lexical causatives is extracted from Levin's (1993) verb index. The verbs that do not alternate were manually removed from the list. All the instances where these verbs occur as transitive, intransitive, and passive were extracted from the automatically parsed English side of the Europarl corpus. We extract the same counts that were extracted for the small sample discussed in Section 4.1.2. We reduce the variance in the corpus-based measure, preserving the information about the ranking of verbs by transforming the frequency ratio into a measure of *Sp* whose range is more limited. We calculate the value of *Sp* (as in 6) for each verb *v* included in the study as the logarithm of the ratio between the proportions of anticausative and causative uses of the verb in the corpus, as shown in (6).

(6)

$$Sp = ln\left(\frac{pr(v, caus)}{pr(v, anticaus)}\right)$$

The proportions of uses of the three extracted constructions are calculated as in (7), where $pr(form) \in \{anticaus, caus, pass\}$ for each verb.

(7)

$$pr(form) = \frac{F(form, v)}{\sum_{form} F(form, v)}$$

The verbs that tend to be used as anticausative will have negative values for the variable *Sp*, those that tend to be used as causative will be represented with positive values, and those that are used equally as anticausative and causative will have values close to zero. The distribution of the *Sp*-value over the 354 verbs is shown in Figure 2 (also later on in Figure 3). For verbs that were not observed in one of the three forms, we calculate the rate values as the rate of uses of the form in the instances of all verbs with frequency 1 divided by the total frequency of the verb in question. For example, the verb *attenuate* occurred three times in the corpus, once as causative and twice as passive. The rate of anticausative uses for this verb is 0.31/3 = 0.10. The number 0.31 that is used instead of the observed count 0 represents the proportion of all verbs with frequency 1 that occurred as intransitive. After normalizing, the proportion of causative uses of this verb is 0.30. The rate of passive uses is 0.61, and the rate of anticausative uses is 0.09. In this manner, we obtain small non-zero values proportional to the two observed frequencies.

4.2.2 Results and discussion

We compare the distribution of the *Sp*-values over the small and the large set of verbs in several ways. Figure 2 illustrates the density distribution of the *Sp*-value over the two samples of verbs.

First, a visual assessment of the shapes of the two distributions suggests that they are similar. Both have a single mode (a single most probable value). Both modes are situated in the same region (around the value 0). The difference in the probability of the most probable values that can be observed in the figure (0.6 for the large sample as opposed to 0.3 for the small sample) does not necessarily reflect the real difference in the two distributions. It can be explained by the fact that the large sample contains numerous unobserved verbs that are assigned the same *Sp*-value, estimated on the basis of the values of low frequency verbs, as discussed earlier. In reality, the verbs would not

have exactly the same value, so that the density would be more equally distributed around zero, which is exactly the case in the small sample.

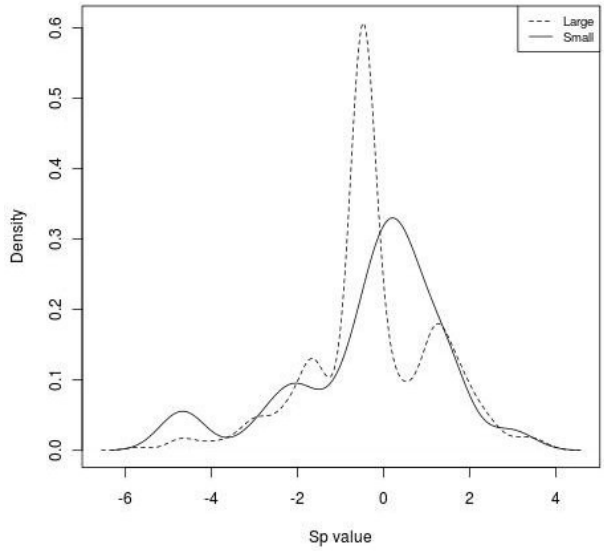

Figure 2: Density distribution of the *Sp*-value in the two samples of verbs

Another indication that the two samples are the same is the value of a two-vector *t*-test $t = -0.0907$, $p = 0.9283$, which indicates a small difference in the means of the two distributions and a high probability that it is observed by chance. The *t*-test works under the assumption that the distributions that are compared belong to the family of normal distributions. It does not apply to other types of distributions. To make sure that the distributions of our data can be compared with the *t*-test, we perform the Shapiro–Wilk test that shows how much a distribution deviates from a normal distribution. This test was not significant ($W = 0.9355$, $p = 0.07$), which means that the distribution of our data can be considered to be normal. We conclude that the verbs for which the corpus-based ranking is shown to correspond to the typology-based ranking represent an unbiased sample of the larger population of verbs that participate in the causative alternation. This implies that the corpus-based method for calculating the *Sp*-value presented in this section can be applied to all the verbs that participate in the alternation.

As a secondary outcome of this experiment, we notice not only that the *Sp*-value is normally distributed over verbs but also that the peak of the distribution is around the value 0. In other words, most verbs are neutral with respect to external causality, whereas only some tend to be assigned extreme values. The shape of the distribution plays an important role in our explanation of the cross-linguistic variation in the alternation availability. In the following section, we take a closer look at the relation between the *Sp*-value and patterns of cross-linguistic variation in the instances of lexical causatives.[8]

4.3 Experiment 3: *Sp*-value and cross-linguistic variation

[8] The verbs selected by Haspelmath (1993) are not deadjectival. They are also a representative sample of the verbs listed by Levin (1993), of which many are deadjectival. This fact means that the distinction between deadjectival and not deadjectival verbs does not play a role in the phenomena studied here.

In analyzing cross-linguistic variation in the realizations of lexical causatives, we determine whether a verb can be expected to have consistent or inconsistent realizations across languages depending on the degree to which an external causer is involved in the event described by the verb.

We approach this task by analyzing German translations of English lexical causatives as they are found in a parallel corpus. Examining instances of translations of lexical causatives in a parallel corpus enables us to control for any pragmatic and contextual factors that may be involved in a particular realization of a lexical causative. Since translations are supposed to express the same meaning in the same context, we can assume that the same factors that influence a particular realization of a verb in a clause in one language influence the realization of its translation in the corresponding clause in another language. Any potential differences in the form of the two parallel clauses should be explained by the lexical properties of the verbs or by structural differences between the languages.

A potential problem with using parallel corpora for linguistic research is posed by translation effects. It has been shown that the language of translated texts differs from the language of original texts in several respects. Baroni and Bernardini (2006) found, for example, that, given a choice between an expression that is similar to the one in the source language and an expression that is different, translators tend to choose the different expression. This tendency produces greater divergence in the translated texts than that imposed by structural differences between the languages. Furthermore, different translators might have different strategies in choosing the expressions. These limitations are partially addressed through the strategy of maximizing parallelism in extracting the instances of verbs, as described below. Using large amounts of data also addresses these issues. A large sample of instances includes translations produced by many different translators so that the individual patterns of variations are cancelled out and the remaining observed variation can be assigned to linguistic factors. The reasoning behind this expectation is that the translator's choice of expression is still limited by linguistic factors: they can only choose between options provided by structural elements available in a language.

We consider English and German, which are genetically and geographically close, a minimal pair. We can expect fewer lexical types to be differently realized in English and German than would be the case in two distant languages, with fewer potential sources of variation. Conversely, if a lexical type is inconsistently used in English and German, inconsistent realizations of this type can be expected in any two more distant languages. This approach is in line with some trends in theoretical linguistics, the so-called study of micro variation in syntax, which attempts to control for large systemic differences by comparing similar languages (Kayne 2005).

Despite the fact that English and German are closely related languages, systematically different realizations of lexical causatives could be expected on the basis of the grammatical and lexical differences that have already been identified in the literature.

It has been noticed that the sets of alternating verbs in these languages are not the same (Schäfer 2009). For example, English verbs of manner of motion, such as *run*, *swim*, *walk*, and *fly* alternate, having both an anticausative and a causative version. Their lexical counterparts in German can only be found as anticausatives (intransitives). The causative use of these verbs in English necessarily requires a transformation in the German translation. At a more general level, it has been claimed that the relations between the elements of the argument structure of German verbs are more specified than those in English verbs, especially in prefixed verbs (Hawkins 1986). Since the specificity of meaning of the verb can influence the alternation, this difference might influence the way in which the verbs are realized in the two languages. From the morphological perspective, the alternation is differently marked in the two languages. While English shows a preference for labile verb pairs – pairs that do not have any morphological marking – German uses both the mechanism of anticausative marking and labile pairs. The availability of different morphological markings might result in differences in how the verbs are used in the two languages.

We perform a statistical analysis of a sample of parallel instances of lexical causatives in English and German, which we divide into three subsamples: expressions of events without external causer, those with external causer, and those that are neutral with respect to the presence of an external causer (low *Sp*-value, high *Sp*-value, and mid-*Sp* value, respectively). This general tendency of the event may or may not be congruent with the actual grammatical context. The two are not congruent,

for example, when verbs that have low *Sp*-value (i.e., that are preferentially used in intransitive contexts) are used in a transitive causative context.

Translations are meaning and structure preserving. If we assume no effect of the *Sp*-value, we can expect a general tendency in favor of parallel translations. If we assume, on the other hand, an effect of the *Sp*-value, we expect syntactic realizations consistent with the lexical semantics of the verb (expressed by its general tendency, hence the *Sp*-value) to be carried across languages in a parallel fashion, while those that are inconsistent are expected to show a tendency toward the consistent realization. For example, we expect intransitive realizations to stay intransitive, and transitives to be often transformed into intransitives when verbs describe low-*Sp* events. Since the probability of both realizations is similar in neutral instances, we expect to find fewer transformations than those in the other two groups.

4.3.1 Materials and methods

The data collected for this analysis come from large and complex resources. We extract the parallel instances from a parallel English-German Europarl (Koehn 2005) corpus, version 3. The corpus comprises German translations of around 1,500,000 English sentences that have been used in the two previous experiments (see Section 4.1). Note that by German translation we mean German translation equivalents since the direction of translation is not known for most of the corpus. The sentences are word-aligned and parsed. Both the syntactic parses and word alignments are provided by Bouma et al. (2010). To collect the information about the syntactic form of the instances needed for our research, the German side of the corpus is syntactically parsed using the same parser as for the English side, the MaltParser (Nivre et al. 2007). The corpus is word-aligned using the system GIZA++ (Och and Ney 2003). We extract the data for our research reusing the processed corpus (with some adaptation and conversion of the format of the data). The verbs included in this study are the 354 English verbs listed as alternating by Levin (1993), for which we have calculated the *Sp*-value applying the procedure described in Section 4.2.

We collect data in the processed parallel corpus from those sentence pairs that have all the following four pieces of information: English syntactic parse, alignment between English and German sentences, alignment between English and German words, and German syntactic parse.[9] We search the English side of the collected sentence pairs to identify English lexical causatives in the same way as in Section 4.2. Extracted instances are represented as pairs or triples of elements. If the instance is intransitive, it is represented with a pair of words: the main verb and the lexical head of the subject. If the instance is transitive, it comprises three elements: the main verb, head of its subject, and head of its object. The German translation of each instance of an English lexical causative is extracted on the basis of word alignments. Instances where at least one German element was word-aligned with at least one element in the English instance were considered to be aligned. The extraction procedure is shown in more detail in Algorithm 2.

We define the alignment in this way to address the issue of missing alignments. The evaluation of the performance of the word-alignment system GIZA++ on the Europarl data for English and German (Pado 2007) showed a recall rate of only 52.9%, while the precision was very high (98.6%).[10] The low recall rate means that around half of the word alignments are not identified by the system. Extracting only the instances where there is a word alignment between an English and a German verb would considerably reduce our dataset. Instead, we rely on the extracted syntactic relations and on the intuition that the verbs are aligned if any of the constituents that depend on them are aligned. With this definition of instance alignment, we also take advantage of our own finding that nouns (the heads of the dependent constituents are nouns) are generally better aligned than verbs (Samardžić and Merlo 2010).

[9] All four pieces of information needed for our study were not available in all the sentences in the Bouma et al. (2010) dataset. To reuse this dataset, it was necessary to find the subset of all the sentences from which all the features could be gathered.

[10] These results apply to an alignment obtained by the intersection of both directions of word alignments (from English to German and from German to English), as is commonly done to improve precision.

Algorithm 2: Extracting parallel cross-linguistic realizations of lexical causatives.

Input: 1. A corpus *E* comprising English sentences *e* that
a. contain realizations of lexical causatives and
b. are parsed with a dependency parser and
c. are annotated with the form of the realization (transitive, intransitive, or passive)

2. A corpus *G* comprising German sentences *g* that are
a. sentence- and word-aligned with the English sentences in *E* and
b. parsed with a dependency parser

Output : Parallel instances comprising
a. the form of the English realization (transitive, intransitive, or passive) and
b. the form of the aligned German realization (transitive, intransitive, or passive)

for $k = 1$ to $k = |E|$ **do**
  if *align*($verb_k$) is a German verb **then**
    g-verb ← *align*($verb_k$);
    do *Algorithm-1(g-verb)*
  else
    if *there is align*(OBJ_k) **then**
      g -verb ← *the verb on which align*(OBJ_k) *depends*;
      do *Algorithm-1(g-verb)*
    else
      if *there is align*($SUBJ_k$) **then**
        g-verb ← *the verb on which align*($SUBJ_k$) *depends*;
        do *Algorithm-1(g-verb)*
      else
        return *no align;*

A pair of extracted aligned instances is illustrated in Table 4. The first column is the sentence identifier; second column is the verb found in the instance; third column is the form of the verb in the instance; and following three columns are the positions of the verb, head of its subject, and head of its object in the sentence.[11]

Applying the method described enables us to extract only translations with limited cross-linguistic variation. Only the instances of English verbs that are translated with a corresponding finite verb form in German are extracted, excluding the cases where English verbs are translated into German with a corresponding non-finite form such as an infinitive, a nominalization, or a participle.

Table 4: Example of an extracted instance of an English alternating verb and its translation to German. The digits under the elements of the realizations of the verbs indicate their position in the sentence. For example, the object of the English verb is the eighth word in the sentence 96-11-14.867, and the object of the German verb is the fifth word in the sentence 96-11-14.859.

[11] One more processing step was needed to identify sentence constituents that are word-aligned because the word alignments and the syntactic analysis did not refer to the same positions. This is caused by the fact that sentence alignment was often not one-to-one. In the cases where more than one English sentence were aligned with a single German sentence, or the other way around, the positions of words were determined with respect to the alignment chunk and not with respect to the individual sentences. For example, if two English sentences were aligned with one German sentence, with eight words in the first sentences and seven in the second, the position of the first word in the second sentence is indicated as 9. In the syntactic parse, on the other hand, these two sentences are not grouped together, so the position of the same word is indicated as 1. We restored the original sentence-based word enumeration in the word alignments before extracting the alignment of the constituents.

Sentence ID	Verb	Form	Verb instance	Subject	Object
96-11-14.867	intensify	CAUS	7	5	8
96-11-14.859	beschleunigen	CAUS	6	2	5

The extracted parallel instances are then combined with the information about the *Sp*-value for each verb to form the final dataset for our study, as illustrated in Table 5. Each line represents one instance of an alternating English verb found in the corpus and its translation into German. The first column contains the *Sp*-value of the verb found in the instance. The second column represents the form in which the English verb is realized in the instance. The third column represents the form of the German translation of the English verb. The full dataset contains 13,033 such items.

Table 5: Examples of parallel instances of lexical causatives.

Sp	En	Ge
1.20	pass	intrans
1.97	trans	trans
0.71	trans	trans
-0.05	pass	pass
0.71	trans	trans
-0.09	trans	pass
-0.14	trans	intrans
-3.91	intrans	intrans
0.39	pass	intrans
-1.76	intrans	trans

Since all the data used in our study are collected automatically from an automatically parsed and word-aligned corpus, they necessarily include processing-related errors. The best reported labeled attachment score of the MaltParser system for English is 88.11 (CoNLL Shared Task 2007) and for German is 85.82 (CoNLL Shared Task 2006). We perform a manual evaluation of a sample of randomly selected instances to assess the degree to which they correspond to the actual analyses. One hundred parallel instances were randomly selected from 13,033 extracted instances: the form of the clause in the English instance and the form of the clause in the German translation were evaluated.

The extraction script assigned an incorrect form to 8/100 English instances (error rate 8%). In seven out of eight errors, the wrong form was assigned due to parsing errors. One error was due to the fact that the parser's output does not include information about traces. For example, in a sentence such as *That is something we must change*, the anticausative form is assigned to the instance of the verb *change* instead of the causative form. In four of the seven parsing errors, the actual forms found in the instances were not verbs but were adjectives or nouns (*open*, *close*, *clear*, *worry*).

The evaluation of the translation extraction was performed only for the cases where the English instance actually contained a verbal form (96 instances). An incorrect form was assigned to the German translation in 13/96 cases (error rate 13.5%). In seven of the thirteen wrong assignments, an incorrect form was assigned to the translation due to parsing errors in German. The errors in three cases were due to the fact that German passive forms headed by the verb *sein* – such as in *Das Team war gespalten,* for English *The team was split* – were not recognized as passives, but they were identified as anticausatives instead. The ambiguity between such forms and anticausative past tenses formed with the *sein* auxiliary verb cannot be resolved in our current extraction method. In the last three cases, the error was due to the fact that the corresponding German form was not a clause. In these cases, the English verb was aligned to a word with a different category (an adverb and a noun) or entirely left out in the German sentence (a verb such as *sit* in *We sit here and produce...*). The

form that was assigned to the translation in these cases was the form of the verb on which the aligned words depend. Our extraction method cannot deal with these cases at the moment, although it would be interesting to capture such mappings.

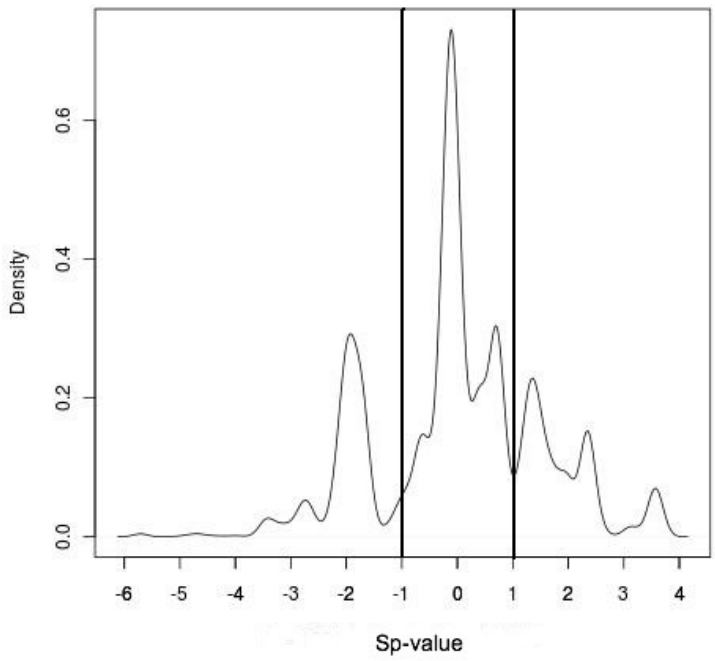

Figure 3: Distribution of all the instances of the 354 verbs (density) over the *Sp*-value.

The validation shows that there could be a bias in our method toward labeling the instances as anticausatives when they are not, especially in German. Final counts seem to be consistent with such a bias (there are more anticausatives than other realizations). However, wrong anticausative counts can be regarded as noise, as they are evenly distributed across groups and do not influence the differences between groups. Three groups of instances are defined according to the density distribution of the *Sp* value. As can be seen in Figure 3, the values -1 and 1 roughly mark symmetric points of low density. We define the instances containing the verbs with the value of *Sp* smaller than -1 as the low-value group. These are expressions of events that do not often require an external causer in terms of the scale of spontaneous occurrence or of internally caused events in the sense of the theories presented in Section 2. Instances containing a verb with the *Sp*-value greater than 1 belong to the high-value group, representing expressions of mostly externally caused events. The instances in between these two values are the medium value instances, representing what Alexiadou (2010) refers to as cause-unspecified events.

This partition gives symmetric subsamples of comparable size: a similar number of examples for the two extreme values (3,107 instances with high *Sp*-values and 2,822 instances with low *Sp*-values) and roughly double this number for non-extreme values (7,104 instances with medium *Sp*-values).

4.3.2 Results and discussion

Table 6 presents the frequencies of the realizations of lexical causatives in parallel English and German instances for the whole sample of instances as well as for the three subsamples. The three most frequent combinations of forms in each group of parallel instances are highlighted to show the changes in the distribution of combinations of forms in the two languages across groups.

The overview of the frequencies suggests that the underlying *Sp*-value of the verb types influences their cross-linguistic realizations. The panel that shows occurrences over the whole

sample indicates that, both in English and in German, intransitives are more frequent than transitives, which are, in turn, more frequent than passives (marginal distributions). Most translations are parallel, and the non-parallel translations cover a third of the cases (32%). When we partition the occurrences by the *Sp*-value of the event, the distribution changes, despite the fact that these are distributions of translations, and are subject to a strong pressure in favor of parallel constructions. Like the distribution of forms over the whole set, the groups of instances containing verbs with a medium *Sp*-value are most often parallel forms, with an even more markedly uniform distribution (29% of non-parallel translations). This means that the verbs that describe events without a strong preference or dispreference for expressing a causer tend to be used in the same form across languages.

Table 6: Contingency tables for the English and German forms in different samples of parallel instances.

Whole sample		German							
		Intransitive		Transitive		Passive		Total	
		N	%	N	%	N	%	N	%
English	Intransitive	3504	**27**	1001	8	314	2	4819	37
	Transitive	1186	9	2792	**21**	369	3	4347	33
	Passive	781	6	517	4	2569	**20**	3867	30
	Total	5471	42	4310	33	3252	25	13033	100
Low *Sp*-value events		German							
		Intransitive		Transitive		Passive		Total	
		N	%	N	%	N	%	N	%
English	Intransitive	1733	**61**	495	**17**	102	4	2330	82
	Transitive	182	**6**	132	5	18	1	332	11
	Passive	35	1	23	1	102	4	160	6
	Total	1950	68	650	23	222	9	2822	100
High *Sp*-value events		German							
		Intransitive		Transitive		Passive		Total	
		N	%	N	%	N	%	N	%
English	Intransitive	74	2	72	2	29	1	175	5
	Transitive	288	9	948	**31**	125	4	1361	44
	Passive	448	**14**	289	10	834	**27**	1571	51
	Total	810	25	1309	43	988	32	3107	100
Mid *Sp*-value events		German							
		Intransitive		Transitive		Passive		Total	
		N	%	N	%	N	%	N	%
English	Intransitive	1697	**24**	434	6	183	3	2314	33
	Transitive	716	10	1712	**24**	226	3	2654	37
	Passive	298	4	205	3	1633	**23**	2136	30
	Total	2711	38	2351	33	2042	29	7104	100

The distribution of the forms is different in the groups of instances containing verbs that describe events at the extremes of the scale of probability of a causer, high or low *Sp*-value. The parallel realizations are frequent only for the forms that are consistent with the lexical properties (intransitive for low *Sp*-value events and transitive for high *Sp*-value events). An atypical instance of a verb in one language (e.g., transitive instance of a verb that describes a low *Sp*-value event) is not preserved across languages. These realizations tend to be transformed into the lexically congruent syntactic

form in the target language. The atypical realizations of these verbs are thus rare across languages, which means that they might be entirely absent in some languages.

This trend is more apparent in Table 7, which shows the probability distributions over German translation forms for a given English form. We can see in the table that the probability that an English form is translated with the same German form is much lower if the English realization is not consistent with the semantic class. The probability that an intransitive form of an English verb is translated with an intransitive German form is 0.73 in the whole sample, 0.74 if the instance is a realization of a low *Sp*-value verb (consistent with the class), 0.73 if it is an instance of a neutral verb, but only 0.42 if the instance is a realization of a high *Sp*-value verb (inconsistent with the class). Simultaneously, the probability of such an instance to be translated with a transitive form is much higher in the last case (0.41) than in all the other cases (0.19–0.21). Exactly the opposite applies to English transitive realizations. The probability of a transitive form in the German translation is the highest if the instance is a realization of a high *Sp*-value verb (0.70), it is still high in the realization of neutral verbs (0.65) and in the whole sample (0.64), but it is much lower in the realizations of low *Sp*-value verbs (0.41). The probability of an English transitive instance to be transformed into an intransitive German form is 0.55 in the last case compared with 0.21–0.28 in the other three cases.

Table 7: Probability distribution over parallel German forms for realizations of English lexical causatives.

Whole sample		Probability of the German translation form			Total	
		Intransitive	Transitive	Passive	N	%
English form	Intransitive	0.73	0.21	0.06	4819	37
	Transitive	0.27	0.64	0.09	4347	33
	Passive	0.20	0.13	0.67	3867	30
Low *Sp*-value events		Probability of the German translation form			Total	
		Intransitive	Transitive	Passive	N	%
English form	Intransitive	0.74	0.21	0.05	2330	82
	Transitive	0.55	0.41	0.04	332	11
	Passive	0.22	0.14	0.64	160	6
High *Sp*-value events		Probability of the German translation form			Total	
		Intransitive	Transitive	Passive	N	%
English form	Intransitive	0.42	0.41	0.17	175	5
	Transitive	0.21	0.70	0.09	1361	44
	Passive	0.29	0.18	0.53	1571	51
Mid *Sp*-value events		Probability of the German translation form			Total	
		Intransitive	Transitive	Passive	N	%
English form	Intransitive	0.73	0.19	0.08	2314	33
	Transitive	0.28	0.65	0.07	2654	37
	Passive	0.14	0.10	0.76	2136	30

We conclude that the analysis of the realizations of lexical causatives in a parallel corpus provides evidence that the probability of occurrence of an external cause in the event described by a verb (the *Sp*-value of the event) is a grammatically relevant lexical property that influences structural realizations. The pattern observed in the cross-linguistic variation can be interpreted as being a consequence of the distribution of the *Sp*-value over verb types established in the previous experiments. The overview of the cross-linguistic joint frequencies enables us to observe directly what this distribution brings to cross-linguistic variation.

Note that the morphological encoding is irrelevant to these findings. The effect concerns only the relation between the underlying property of a verb type and its syntactic realizations across languages. The account proposed in the study by Haspelmath et al. (2014) cannot address the pattern established in this experiment since that approach explains the cross-linguistic morphological

encoding patterns through a principle of efficiency. The underlying property of *Sp*-value, assigned to verb types, accounts for both morphological and syntactic patterns.

Accounting for the asymmetry in cross-linguistic realization of the alternating verbs is important because this pattern suggests a probabilistic explanation of why some languages exhibit the causative alternation for a given verb and some languages do not. Verbs that describe events at the extremes of the distribution of *Sp*-value (left and right tail) are more likely to have different realizations across languages than those that describe events in the middle of the scale (around the peak). When the probabilities of the two realizations are similar, in verbs with mid *Sp*-value, these realizations can be expected to occur with similar frequency across languages. Since both realizations are frequent in these verbs, they can be expected to alternate in the majority of languages. When the *Sp*-values are at the low and high end of the scale, the atypical realizations of these verbs are rare across languages, which means that they might be entirely absent in some languages. This means that the verbs describing events at the extremes of the scale of *Sp*-values can be expected to alternate in a smaller range of languages than verbs in the middle of the scale. The distribution of the *Sp*-value across verb types does not predict which verb types alternates in which particular language, but it predicts in how many languages it will alternate for each verb type.

4.4 Experiment 4: Learning spontaneity with a probabilistic model

Having showed that the probability of external causation can be estimated from the distribution of causative and anticausative instances of verbs in a corpus (sections 4.1 and 4.2) and that this continuous feature affects cross-linguistic realizations of lexical causatives (Section 4.3), we now turn to the issue of identifying classes of lexical causatives.

In contrast to all previous approaches reviewed in Section 2, we address the issue of classification of lexical causatives in a fully empirical and automatic fashion. Building on the findings of the first three experiments, we train a Bayesian probabilistic model that takes as input the corpus frequency of anticausative and causative uses of verbs and generates a probability distribution over a number of possible classes for each verb type. The model is trained in an unsupervised learning framework: no predefined classes are encoded in the training data; the model infers the classes by observing patterns in the distribution of the verb realizations. The model only assumes that an *Sp* variable representing the lexical semantic class exists, and it automatically infers if the observed distributional corpus data are better explained by two or three classes.

By varying the settings of the model, we address two questions discussed in the linguistic literature. First, we ask if the distinction between externally caused and internally caused events is binary, as argued by Levin and Rappaport Hovav (1994), or if there are intermediate classes, as argued by Alexiadou (2010). Then, we ask if better estimates of the classes are obtained from cross-linguistic or from monolingual data.

Recall that two main proposals concerning the classification of alternating verbs have been proposed in the linguistic literature. As discussed in Section 2, Levin and Rappaport Hovav (1994) used the distinction between externally caused and internally caused events to explain a set of observations concerning the alternating verbs. Alexiadou (2010), however, indicated that a range of cross-linguistic phenomena are better explained by introducing a third semantic class, the cause-unspecified verbs. The distinctions argued for in the linguistic literature can be quantified by our *Sp*-value, which encodes if events are likely to have an external causer or not. The classes can be established by partitioning the range of *Sp*-values found in the previous experiments into discrete intervals. We experiment with two proposed classifications, but the same approach can be extended to any number of hypothesized classes.

Various empirical studies have shown that languages have different biases that can influence the realization of lexical causatives. For example, a quantitative typological study by Nichols et al. (2004) indicated that some languages have a preference for transitivization of certain lexical types. The findings of Haspelmath et al. (2014) indicated that the use of verbs in the corpora of all seven individual languages included in the study deviates from the general tendencies. The agreement between the corpus-based and typology-based ranking of verbs is better when the counts are

averaged over the languages than for any individual language. Finally, studies have also established that English prefers causative realizations of some verbs compared with other languages (Bowerman and Croft 2008; Wolff et al. 2009a; Wolff and Ventura 2009b). To address the issue of the potential influence of language-specific factors on corpus-based identification of the classes of lexical causatives, we extend our corpus-based model of verb classes to the cross-linguistic domain. We collect the information about the realizations of the alternating verbs in a parallel corpus, as described in Section 4.3. The extended dataset is expected to provide a better estimation of the *Sp*-value than the monolingual set, neutralizing language-specific influences. Naturally, including more languages would be expected to give even better estimates. In this study, however, we consider only two languages as a first step toward a richer cross-linguistic setting.

To address the two issues mentioned above – number of underlying lexical classes and influence of language-specific biases – we design four experimental settings, varying the input data and the number of classes in the output: a) monolingual input and two classes; b) cross-linguistic input and two classes; c) monolingual input and three classes; d) cross-linguistic input and three classes.

4.4.1 The model

As can be seen in its graphical representation in Figure 4, the model comprises three variables in the monolingual version and four variables in the cross-linguistic version. The first variable is the set of considered verbs *V*. This can be any given set of verbs. The second variable is the *Sp*-class of the verb, for which we use the symbol *Sp*. The values of this variable depend on the assumed classification. The third (*En*) and fourth (in the bilingual version, *Ge*) variables are the surface realizations of the verbs in parallel instances. These variables take three values: causative for an active transitive use, anticausative for an intransitive use, and passive for a passive use.

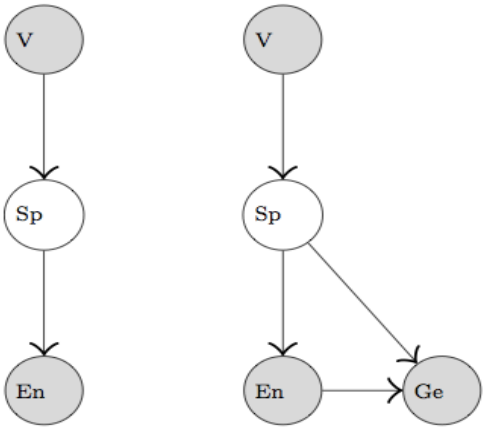

Figure 4: Two versions of the Bayesian net model for estimating external causation.

We represent the relations between the variables by constructing a Bayesian network, as illustrated in Figure 4. A Bayesian network is a directed acyclic graph where the nodes represent the variables of a model and edges represent the dependency relations between the variables (Russell and Norvig 2010). The variable that represents the *Sp*-class of verbs is treated as an unobserved variable. The values for the other three variables are observed in the data source. Note that the input to the model, unlike the information extracted for the analysis in Section 4.3, does not contain the information about the *Sp*-value (compare Table 8 with Table 5). The dependence between *En* and *Ge* represents the fact that the two instances of a verb are translations of each other but does not represent the direction of translation. The form of the instance in one language depends on the form of the parallel instance because they express the same meaning in the same context, regardless of the direction of translation.

Table 8: Examples of the crosslinguistic input data

V	En	Ge
move	pass	intrans
alter	trans	trans
improve	trans	trans
increase	pass	pass
improve	trans	trans
break	trans	pass
change	trans	intrans
grow	intrans	intrans
close	pass	intrans
split	intrans	trans

Assuming that the variables are related to each other as in Figure 4, *En* and *Ge* are conditionally independent of *V* given *Sp*; thus, we can calculate the probability of the model as in (8) for the monolingual version and as in (9) for the cross-linguistic version.

(8) $$P(v, sp, en) = P(v) \cdot P(sp|v) \cdot P(en|sp)$$

(9) $$P(v, sp, en, ge) = P(\ v) \cdot P(sp|v) \cdot P(en|sp) \cdot P(ge|sp, en)$$

We estimate the conditional probability of the event class given the verb $P(sp|v)$ by querying the model, as shown in (10) for the monolingual version and in (11) for the bilingual version.

(10) $$P(sp|v, en) = \frac{P(v) \cdot P(sp|v) \cdot p(en|sp)}{\sum_{sp} P(v) \cdot P(sp|v) \cdot P(en|sp)}$$

(11)
$$P(sp|v, en, ge) = \frac{P(\ v) \cdot P(sp|v) \cdot P(en|sp) \cdot P(ge|sp, en)}{\sum_{sp} P(\ v) \cdot P(sp|v) \cdot P(en|sp) \cdot P(ge|sp, en)}$$

Having estimated the *Sp*-value for each verb instance, we assign to each verb the average *Sp*-value across instances, as shown in (12), where *F(v)* denotes the number of occurrences of the verb in the training data.

(12)
$$sp\_class(verb) = \frac{\sum_{en} \sum_{ge} P(sp|v, en, ge)}{F(v)}$$

All the variables in the model are defined so that the parameters (the different probabilities in the model) can be estimated on the basis of frequencies of instances of verbs automatically extracted from parsed corpora. The corpus used as input does not need to be annotated with classes since the parameters are estimated treating the class variable as unobserved.

We estimate the parameters of the model by implementing the expectation-maximization algorithm (Dempster et al. 1977), initialized randomly. The classification reported in the paper is obtained after 100 iterations.

4.4.2 Evaluation and discussion

We implement a classifier based on the model, which we train using the data extracted from the syntactically parsed and word-aligned parallel corpus of English and German, as described in Section 4.3. The verbs used to estimate the model are the 354 verbs that participate in the causative alternation in English, as listed by Levin (1993). For the set of verbs for which the typological information is available, we compare the classification of verbs learned by the models using the typology-based ranking. To evaluate the output of our models against the typological scale, proposed independently by Haspelmath (1993), we discretize the scale so that the agreement is maximized for each version of the model.

Table 9: Agreement between corpus-based and typology-based classification of verbs (a = anticausative, c = causative, m = cause-unspecified).

Verb	Two-way classification		Three-way classification	
	Monolingual	Bilingual	Monolingual	Bilingual
boil	a	a	a	a
dry	a	a	a	a
wake up	a	a	a	a
sink	a	a	a	a
learn-teach	a	a	m	m
melt	a	a	a	a
stop	a	a	a	a
turn	a	a	m	m
dissolve	c	c	c	c
burn	c	c	c	m
fill	c	c	c	m
finish	a	a	a	a
begin	a	a	a	a
spread	a	a	a	a
roll	c	c	m	m
develop	c	c	m	m
rise-raise	c	c	c	c
improve	c	c	m	m
rock	c	c	m	m
connect	c	c	c	c
change	a	a	a	a
gather	c	c	m	m
open	c	c	m	m
break	c	c	c	c
close	c	c	c	c
split	c	c	c	c
Agreement	85%	85%	61%	69%

Table 9 indicates the output of the four models and the manner in which the scale is discretized. The verbs in the table are ordered according to the typological scale, in decreasing order of morphological coding tendencies (Haspelmath 1993, Haspelmath et al. 2014). The thresholds that maximize agreement between the typological scale and output of the models are represented with

horizontal lines. The outputs of the models are represented with single letters (*a* for the anticausative, *m* for the medium, corresponding to cause-unspecified, *c* for the causative class). For the models to agree with the typological scale, the verbs at the top need to be classified as anticausative, those at the bottom need to be classified as causative, and those in the middle need to be classified as medium.

Table 10: Overall and by-class agreement between the corpus-based and typology-based classification of verbs.

Models / Typology	Two-way classification				Three-way classification					
	Monolingual		Bilingual		Monolingual			Bilingual		
	Antic	Caus	Antic	Caus	Antic	Caus	Mid	Antic	Caus	Mid
Low Sp	8	0	8	0	6	0	1	6	0	1
High Sp	4	14	4	14	0	3	0	0	3	0
Mid Sp	--	--	--	--	4	5	7	4	3	9
Agreement	85%		85%		61%			69%		

The classification performed by the bilingual model indicates that the distinction between externally caused and cause-unspecified verbs might still exist. Compared with the monolingual classification, more verbs are classified as cause-unspecified, and they are grouped in the middle of the typological scale. Since the model considers cross-linguistic variation in the realizations of the alternating verbs, the observed difference in the performance between the monolingual and multilingual models can be interpreted as a sign that the distinction between cause-unspecified and externally caused events emerges in cross-linguistic contexts.

While supporting the two-way classification of events, our experiments do not provide a definitive answer to the question of whether there are more than two classes of events. To obtain significant results, more verbs need to be evaluated. However, the typological data used in our experiments, from the study by Haspelmath (1993), are not easily available. This kind of data are currently not included in typological resources (such as the WALS database, Dryer and Haspelmath [2013]), but they can, in principle, be collected from other electronic sources of language documentation, which are increasingly available for many different languages.

The design of our cross-linguistic model is fundamentally different from those of all the other models with regard to the relation between corpus data and typological tendencies (Haspelmath et al. 2104). The input corpus data in all the other models are either monolingual or aggregated multilingual. Instead, our model is based on parallel data: it keeps the information of which realization comes from which language and which token. In this manner, we are able to account for language-specific biases and potential within-language regularities.

Overall, these experiments support a unified explanation of cross-linguistic and frequency variation in causative alternations in terms of a single gradable property of verb types, namely, the probability of external causation.

5 General discussion

The experiments performed in our study of morpho-syntactic realizations of lexical causatives relate various factors involved in the causative alternation in a systematic fashion based on the single notion of the probability of external causation. The results of all our experiments suggest that the alternating verbs are spread according to their probability of external causation. The probability of external causation (the term *Sp*-value is used in our experiments) is distributed normally over a large number of verb types and verb instances. Since the normal distribution is symmetric and concentrated around the mean, this normal distribution implies that the most likely value is the mean value and that both extremely low and extremely high values are equally likely.

This finding suggests a typological prediction. Since the distribution of the probability of external causation is normal, most alternating verbs can be expected to describe events in which there is

around 50% chance for an external causer to occur. These verbs alternate in the majority of languages. However, the probability of an external causer can be small for some verbs. These verbs alternate only in some languages, while they do not alternate in the majority of languages. The same can be claimed for verbs describing events with extremely likely external causers. Thus, the probability of external causation as a property of verb types explains the observed cross-linguistic variation in the availability of causative alternation. The number of languages in which a verb can be expected not to alternate can be predicted from the probability of external causation assigned to it.

Although it might be a continuum, the scale of probability of external causation can be represented by discrete values. The results of our experiments on classification indeed suggest that clustered values in the discretized scale correspond to the classes proposed in the literature.

As already indicated earlier, two recent studies have argued against a dependency between the lexical properties of verbs and their observable morphology. We now turn to the differences between our view and these two studies. Alexiadou et al. (2015) found no underlying pattern in the distribution of marked, unmarked, and optionally marked anticausatives in Greek, Italian, French, and German. They, however, did not consider the whole range of verbs but considered a relatively small sample of examples. Our study makes a different prediction concerning their data and suggests that there should be a pattern in each language: the unmarked verbs that are classified as anticausative would tend to be placed on the (anticausative) extreme of the scale, those that are consistently marked would be placed closer to its middle, and those optionally marked would lie between the two. The absolute position depends on the language, but the relative position is not. In other words, some languages can start marking the anticausatives only in the intervals that are closer to the middle of the scale, while others can start already close to the anticausative end of the scale. However, once the marking appears, it should not disappear again as we move toward the other end of the scale. This pattern cannot be expected to apply with perfect regularity, but a general tendency (with some random variation) would be predicted to appear in a larger sample. To understand whether there is an underlying pattern in the distribution of the morphological types within a language, one would need to test for such contingency conditions. The methods proposed in our study enable us to locate a large number of verbs in different languages on the scale and empirically test such hypotheses.

Preliminary support for our prediction comes from the results of Heidinger (2015) on French and Spanish anticausatives. Recall that Heidinger measures the notion of causalness (see the end of section 2 above) in the same way as we measure our property of likelihood of external causation, as the ratio of causative and anticausative uses of a verb. Heidinger (2015) also counted how often causatives are overtly marked and how often anticausatives are overtly marked. Heidinger (2015) showed that the two measures (percentage of causative marking and causalness) correlate. Interestingly, however, the data distribution for percentage of markedness is a step function (a function that abruptly changes value, creating a "step," instead of the usual more gradual change creating a slope). The percentage of marked forms is almost discretely distributed: anticausatives that exhibit low-to-mid causalness almost never mark the anticausative form, whereas mid-to high causalness anticausatives almost always do. There is some variation in the causative marking only for those verbs that have around 50% causalness. This is true for both French and Spanish (although Spanish has more variation than French in the marking of low-to mid verbs).

Haspelmath et al. (2014) found an underlying pattern in the distribution of morphological marking similar to the one presented in our study. They, however, gave it a different interpretation. In their view, the observed pattern is a direct consequence of efficient coding for efficient communication: unmarked forms are unmarked because they are frequently used (measured as corpus frequency) regardless of the lexical properties of verbs. In contrast, our interpretation postulates the existence of a "hidden variable," an underlying component anchored in each lexical entry that determines both the corpus frequency distribution and typological distribution of the morphology. While coding efficiency certainly plays a role in the distribution of the morphological marking, it cannot be the only explanation for the results presented by Haspelmath et al. (2014). The association between morphology and frequency is the strongest when the frequency data from seven languages are aggregated, while it is weaker in the case of any individual language (and not confirmed for English). This means that the coding is not optimized for efficiency within a language.

It is unclear why the efficiency principle would apply only globally and not within an individual language, given that the individual languages are the actual means of communication. Moreover, a global efficiency optimization is not possible without a cross-linguistic mediator, such as the *Sp*-value that we propose. If the relation between the frequency and the morphological coding were direct, as argued by Haspelmath et al. (2014), language-specific associations patterns would have to be stronger than or at least as strong as a single global association.

While our main goal was to address the cross-linguistic variation in the availability of the causative alternation, the results of our experiments are also potentially relevant to the much-discussed question of which of the two alternants is basic and which one is derived (Reinhart 2002; Chierchia 2004; Levin and Rappaport Hovav 1995; Rappaport Hovav and Levin 2012, Koontz-Garboden 2009; Rappaport Hovav 2014; Piñón 2001; Alexiadou et al. 2006; Schafer 2008; Alexiadou and Doron 2012). The shape of the distribution of the *Sp*-value and the effects related to it provide empirical support for the view that the direction of the derivation is not the same for all the verbs, but depends on the verb type, as occasionally evoked in accounting for some specific phenomena. Events whose external causer is likely give rise to basic causative verbs. Events whose external causer can be present or absent with similar probability (the great majority of events) generate a neutral basic form, and the events that are unlikely to have an external causer give rise to a basically anticausative verb. This relation is blurred by language-specific biases in syntactic realizations and morphological marking, but it enables the cross-linguistic patterns observed in our experiments to emerge.

The detailed computational analysis proposed in this study enables us to consider a large amount of data. This, in turn, makes apparent some patterns that are not observable in the small samples that are studied in all the other studies, even the empirical ones. For instance, the findings of experiment 4 suggest that the anticausative part of the scale seems to be more easily distinguishable than the other intervals of the scale. The verbs classified as anticausative in our experiments can be related to verbs describing the internally caused events discussed in the literature. Relating these two categories, however, requires a relaxed notion of internal causation (allowing the internally caused verbs to alternate), as it is used in more recent accounts (Levin 2009, Alexiadou 2014). The anticausative class of verbs includes both the verbs that do and those that do not alternate in English, but all these verbs can be expected to alternate in fewer languages than those in the middle of the scale. The question of whether there is a difference between the class of cause-unspecified and causative verbs remains open. If it turns out that these two classes cannot be distinguished, this raises the question of why the two extremes of the scale behave differently from each other with respect to classification. The data collected in our experiments do not seem to provide enough empirical evidence to address these issues. Although some tendencies seem to emerge in our classification experiments, more data from more languages would need to be analyzed before some answers to these questions can be offered.

6 Conclusion

The experiments presented in this study provide new empirical evidence, gathered using a large-scale computational approach. This new evidence clarifies the relation between the lexical properties of the verbs and various aspects of cross-linguistic variation in the realizations of the causative alternation. We have shown that the frequency of use of lexical causatives in corpora is a good predictor of the morphological marking that these verbs receive across languages and of the cross-linguistic spread of the alternation.

An analysis of a large number of instances automatically extracted from parallel corpora enables us to directly observe the influence of the notion of probability of external causation on the cross-linguistic variation in the realizations of lexical causatives. The investigation of two closely related languages, English and German, indicates that the realizations that are consistent with the verb's lexical type tend to be preserved, whereas the inconsistent realizations tend to be transformed across languages.

Finally, the use of large sets of multilingual data enables us to apply computational methods to directly induce classes of lexical causatives from basic observations, considering many different

factors discussed in the literature. Our probabilistic model, which successfully classifies verbs into classes of external causation, learns the classification in an unsupervised fashion, with no predefined categories. The model and the evaluation scheme that we have designed make some specific prediction regarding the interaction between the underlying, unobserved properties of verb types and the observed properties of their form and use. It can be considered as a basis for future experimentation, which will address the language-specific regularities involved in the many realizations of the causative alternation across languages more systematically.

References

Aboh, Enoch. 2009. Clause structure and verb series. *Linguistic Inquiry* 40(1). 1–33.

Alexiadou, Artemis. 2006. *On (anti-)causative alternations*, University of Stuttgart. Presentation, Ecole d'automne de linguistique, Paris.

Alexiadou, Artemis. 2010. On the morpho-syntax of (anti-)causative verbs. In Malka Rappaport Hovav, Edit Doron & Ivy Sichel (eds.), *Syntax, lexical semantics and event structure*, 177–203. Oxford: Oxford University Press.

Alexiadou, Artemis. 2014. The problem with internally caused change-of-state verbs. *Linguistics* 52 (4). 879–909.

Alexiadou, Artemis & Edit Doron. 2012. The syntactic construction of two non-active voices: passive and middle. *Journal of Linguistics* 48. 1–34.

Alexiadou, Artemis, Elena Anagnostopoulou & Florian Schäfer. 2006. The properties of anticausatives crosslinguistically. In Mara Frascarelli (ed.), *Phases of interpretation*, 187–212. Berlin, New York: Mouton de Gruyter.

Alexiadou, Artemis, Elena Anagnostopoulou & Florian Schäfer. 2015. *External arguments in transitivity alternations: A layering approach* (Oxford studies in Theoretical Linguistics). Oxford: Oxford University Press.

Alsina, Alex. 1997. A theory of complex predicates: Evidence from causatives in Bantu and Romance. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex predicates*, 203–247. Stanford, California: CSLI Publications.

Baker, Mark C. 1988. *Incorporation: A theory of grammatical function changing*. Chicago & London: The University of Chicago Press.

Baroni, Marco & Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing,* 21(3). 259–274.

Bouma, Gerlof, Lilja Øvrelid & Jonas Kuhn. 2010. Towards a large parallel corpus of cleft constructions. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10),* Valletta: European Language Resources Association.

Bowerman, Melissa & William Croft. 2008. The acquisition of the English causative alternation. In Melissa Bowerman & Penelope Brown (eds.), *Crosslinguistic perspectives on argument structure: Implications for learnability*, 279–306. New York, NY: Lawrence Erlbaum.

Chierchia, Gennaro. 2004. A semantics for unaccusatives and its syntactic consequences. In Artemis Alexiadou, Elena Anagnostopoulou & Martin Everaert (eds.), *The unaccusativity puzzle*, 22–59. Oxford: Oxford University Press.

Collins, Chris. 1997. Argument sharing in serial verb constructions. *Linguistic Inquiry* 28. 461–497.

Croft, William. 1990. Possible verbs and the structure of events. In Savas L. Tsohatzidis (ed.), *Meanings and prototypes*, 48–73. London: Routledge.

Dempster, Arthur P., Nan M. Laird & Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39 (1). 1–38.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Hale, Ken & Jay Keyser. 1993. On argument structure and the lexical representation of syntactic relations. In Ken Hale & Jay Keyser (eds.), *The View from Building 20,* 53–110. Cambridge, Mass: MIT Press.

Haspelmath, Martin. 1993. More on typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity* (Volume 23), 87–121. Amsterdam/Philadelphia: John Benjamins.

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19 (1). 1–33.

Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog & Elif Bamyac. 2014. Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics* 50 (3). 587–625.

Heidinger, Steffen. 2015. Causalness and the encoding of the causative–anticausative alternation in French and Spanish. *Journal of Linguistics* 51(3). 562–594.
Hawkins, John A. 1986. *A comparative typology of English and German: Unifying the contrasts*. London & Sydney: Croom Helm.
Horvath, Julia & Tal Siloni. 2011. Anticausatives: against reflexivization. *Lingua* 121. 2176–2186.
Kallulli, Dalina. 2006. A unified analysis of passives, anticausatives and reflexives. In Olivier Bonami & Patricia Cabredo Hofherr (eds.), *Empirical issues in formal syntax and semantics*, 271–301. Amsterdam: John Benjamins.
Kayne, Richard. 2005. Some notes on comparative syntax, with special reference to English and French. In *The Oxford Handbook of Comparative Syntax*. Oxford University Press.
Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT summit 2005*. Phuket.
Koontz-Garboden, Andrew. 2009. Anticausativization. *Natural Language & Linguistic Theory* 27 (1). 77–138.
Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: The University of Chicago Press.
Levin, Beth. 2009. Further explorations of the landscape of causation: Comments on the paper by Alexiadou and Anagnostopoulou. In *Proceedings of the workshop on Greek syntax and semantics* (MIT Working Papers in Linguistics 49), 239–266. Cambridge, Mass.: MIT Department of Linguistics and Philosophy.
Levin, Beth & Malka Rappaport Hovav. 1994. A preliminary analysis of causative verbs in English. *Lingua* 92. 35–77.
Levin, Beth & Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface*. Cambridge, Mass.: MIT Press.
Levin, Beth & Malka Rappaport Hovav. 2013. Lexicalized meaning and manner/result complementarity, in Boban Arsenijević, Berit Gehrke & Rafael Marín (eds.), *Subatomic semantics of event predicates,* 49–70. Springer: Dordrecht.
Marelj, Marijana. 2004. *Middles and argument structure across languages*. Utrecht: LOT.
Merlo, Paola & Susanne Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics* 27 (3). 373–408.
Nichols, Johanna, David Peterson & Jonathan Barnes. 2004. Transitivizing and detransitivizing languages. *Linguistic Typology* 8. 149–212.
Nivre, Joakim, Johan Hall, Jens Nilsson, Chanev Atanas, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13 (2). 95–135.
Och, Franz Josef & Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29 (1). 19–52.
Pado, Sebastian. 2007. *Cross-lingual annotation projection models for role-semantic information*. PhD diss, Saarland University.
Palmer, Martha, Daniel Gildea & Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31 (1). 71–105.
Piñón, Christopher. 2001. A finer look at the causative-inchoative alternation. In Brendan Jackson, Rachel Hastings and Zsofia Zvolenszky (eds.), *Proceedings of semantics and linguistic theory*, (Volume 11), 346–364. Ithaca, NY: CLC Publications.
Rappaport Hovav, Malka. 2014. Lexical content and context: The causative alternation in English revisited. *Lingua* 141. 8–29.
Rappaport Hovav, Malka & Beth Levin. 2012. Lexicon uniformity and the causative alternation. In Martin Everaert, Marijana Mareli & Tal Siloni (eds.), *The theta system*, 150–176. Oxford University Press.
Reinhart, Tanya. 2002. The theta system – An overview. *Theoretical linguistics* 28. 229–290.
Reinhart, Tanya & Tal Siloni. 2004. Against the unaccusative analysis of reflexives. In Artemis Alexiadou, Elena Anagnostopoulou & Martin Everaert (eds.), *The unaccusativity puzzle: Studies on the syntax-lexicon interface*, 159–181. Oxford University Press.

Russell, Stuart J. & Peter Norvig. 2010. *Artificial intelligence: A modern approach*. Upper Saddle River, N.J.: Prentice Hall Pearson.

Schäfer, Florian. 2008. *The syntax of (anti-)causatives: External arguments in change-of-state contexts*. Amsterdam & Philadelphia: John Benjamins.

Schäfer, Florian. 2009. The causative alternation. In *Language and linguistics compass* (Volume 3), 641–681. Blackwell Publishing.

Samardžić, Tanja & Paola Merlo. 2010. Cross-linguistic variation of light verb constructions: using parallel corpora and automatic alignment for linguistic research. In *Proceedings of the workshop of the ACL workshop NLP and Linguistics: finding the common ground,* 52–60. Uppsala: Association for Computational Linguistics.

Samardžić, Tanja & Paola Merlo. 2012. The meaning of lexical causatives in cross-linguistic variation. *Linguistic Issues in Language Technology* 7 (12). 1–14.

Samardžić, Tanja & Paola Merlo. 2014. Likelihood of external causation in the structure of events, In *Proceedings of the EACL Workshop on Computational Approaches to Causality in Language*, 40-47. Gothenburg: Association for Computational Linguistics.

Stevenson, Suzanne & Paola Merlo. 1997. Lexical structure and parsing complexity. *Language and Cognitive Processes* 12 (2/3). 349–399.

Williams, Edwin. 1997. Lexical and syntactic complex predicates. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex predicates*, 13–29. Stanford, California: CSLI Publications.

Wolff, Phillip & Tatyana Ventura. 2009b. When Russians learn English: How the semantics of causation may change. *Bilingualism: Language and Cognition* 12 (2). 153–176.

Wolff, Phillip, Ga-Hyun Jeon & Yu Li. 2009a. Causal agents in English, Korean and Chinese: The role of internal and external causation. *Language and Cognition* 12 (2). 165–194.