Measuring iconicity:

A quantitative study of lexical and analytic causatives in British English[*]

Abstract

The idea of isomorphism of form and meaning has played an important role in functionalist theories of syntax and morphology. However, there have been few studies that test this hypothesis empirically on quantitative data. This study aims to fill in this gap by testing the predictions made by iconicity theory with the help of statistical hypothesis-testing techniques. The paper focuses on a subtype of isomorphism, namely, iconicity of cohesion. The analyses are based on a sample of lexical and analytic causatives from the British National Corpus. The study employs three different operationalisations of the degree of semantic cohesion of the causing and caused events, which are based on English and cross-linguistic data. The form-function correlation is interpreted from the point of view of three possible models of relationships between form, function and/or frequency.

1. Introduction

The idea of isomorphism of linguistic form and function is well-established in functional and cognitive linguistics. This is not surprising, since both frameworks emerged as a reaction to structuralism and its late manifestation, generative linguistics, which placed emphasis on the arbitrary nature of language. Iconicity has played an important role in functionally oriented syntax and morphology. A well-known example is Givón's (1980) binding hierarchy, which

[*] The author is very grateful to Martin Hilpert, the anonymous reviewers and Martin Haspelmath for their excellent suggestions and constructive criticisms, which have helped to improve the paper considerably. The research is a part of the project financially supported by the Belgian research foundation F.R.S.-FNRS. All usual disclaimers apply.

posits a correlation between the degree of semantic integration of events and syntactic integration of the corresponding predicates. However, to our best knowledge, few quantitative analyses have been performed in order to test whether iconicity-based predictions in fact hold when confronted with real data. Some of the notable exceptions are Bybee (1985a), Rosenbach (2003), Diessel (2008) and Steger and Schneider (2012). A possible reason is a lack of communication across the boundaries of linguistic sub-fields. There is a deplorable gap between quantitative linguists, many of whom are not particularly interested in testing global theories, and functional linguists and typologists, who are usually satisfied with qualitative analyses of individual examples (but see Bickel 2010 and Haspelmath *et al.* 2014 as counterexamples). Another reason is methodological: the correlation between form and function is difficult to test because an operationalisation of functional properties for a quantitative study is by no means obvious. There is a danger of circularity when semantics is inferred from syntax, which, in its turn, is explained by semantics.

Importantly, iconicity theory has been recently challenged by Haspelmath (2008b, Haspelmath *et al.* 2014, and other works). He claims that formal asymmetries are better explained by frequency asymmetries than by iconic relationships (see more details in Section 7). However, since the iconicity explanation has never been tested empirically in a systematic way, this new theory might be trying to defeat an illusory opponent. It may well be that the previous researchers' conclusions based on introspection and mostly self-invented examples were simply erroneous.

The present paper aims to fill in this gap by investigating the division of labour between English analytic and lexical causatives, e.g. *break*$_{\text{TR}}$ and c*ause/make/have/get* (*to*) *break*$_{\text{INTR}}$. We test the predictions made by iconicity theory by measuring the correlation between the ratio of analytic vs. lexical causatives in a large corpus and the degree of integration of the causing and caused events expressed by the causatives. To avoid circularity, we use three context-independent variables that represent the degree of integration of the

causing and caused events. These variables are based on the generic properties of the caused events expressed by the causative constructions, rather than on the characteristics of particular causative situations. Two of these variables are language-internal and one is cross-linguistic. The first language-internal variable is based on a semantic classification of causation events. The second is a fully corpus-driven measure, a ratio of causal and non-causal uses of a verb in a large corpus. The third is a cross-linguistic variable that contains the ratios of anticausative and causative alternations in a sample of languages from Haspelmath (1993). All of these variables represent the degree of spontaneity of the caused event, which is closely related to the degree of semantic integration of events.

This paper employs statistical hypothesis-testing techniques (correlation analysis and logistic regression, along with some other methods) in order to test the correlations between the above-mentioned variables and the ratios of the lexical and analytic causatives for 62 pairs of lexical and analytic causatives. The frequencies of the lexical and analytic causatives for each pair are extracted from the entire British National Corpus (2007).

The remaining part of the paper is organised as follows. Section 2 offers a discussion of iconicity phenomena in morphosyntax, with a focus on variation of causative constructions. Section 3 gives a definition of analytic and lexical causatives, employing the notion of a comparative concept (Haspelmath 2010), and describes the sample of English analytic and lexical causatives from the British National Corpus. Sections 4 to 6 present the quantitative analyses where the iconicity-based predictions are tested with the help of three different operationalisations. Section 7 provides a discussion of the findings and introduces three possible theoretical models that may explain the variation. Finally, Section 8 offers some concluding remarks.

2. Iconicity theory and causatives

According to Givón (1990: 968–973), the main manifestations of iconicity in syntax are the following:

- the Quantity principle: a larger chunk of information will be given more coding material. The same holds for less predictable or more important information. An example is the larger size and more prominent stress of lexical words in comparison with grammatical morphemes;

- the Proximity principle: entities that are closer together functionally, conceptually or cognitively tend to be placed closer together formally (temporally, in spoken language, or spatially, in written language). Examples are Givón's (1980) binding hierarchy and variation in causative constructions, which is the focus of the present paper;

- Sequential order principles, which include the principles of sequential order and topicality. According to the former, the order of clauses will tend to correspond to the temporal order of the occurrence of events depicted in the discourse (cf. Diessel 2008). For example, there is a strong tendency to place the conditional clauses before the clause that contains the entailment (*if P, then Q*). As for topicality, one can expect highly important or urgent information to be placed first in the string. For example, contrastive topics are usually placed in the clause-initial position, as in *it*-clefts, e.g. *It's John (not Bill) who broke the window*.

It was pointed out by Haspelmath (2008b) that the Proximity principle in fact conflates two different phenomena, which can be called iconicity of contiguity and iconicity of cohesion. Iconicity of contiguity means that elements that belong together semantically also occur together. As for iconicity of cohesion, it can be expressed as follows:

(1) “Meanings that belong together more closely semantically are expressed by more cohesive forms” (Haspelmath 2008b: 2).

Formal cohesion is the degree of interdependence of two elements. It is inversely related to formal distance. For example, Haiman (1983: 782) charts a cline of linguistic distance, which is shown in (2).

(2) a. X # A # Y

b. X # Y

c. X + Y

d. Z

In this cline, X, A and Y are morphemes, # represents a word boundary, + stands for a morpheme boundary, and Z is a morpheme where X and Y are fused. The formal distance between morphemes X and Y decreases from (a) to (d). Conversely, one can say that the degree of cohesion of X and Y increases from (a) to (d).

Semantic cohesion is more difficult to define and depends on the type of the linguistic construction discussed. In particular, when speaking about verbal complementation, researchers have proposed such semantic parameters as spatiotemporal integration of events, referential cohesion, control and autonomy of event participants and implicative relationships between propositions (e.g. Givón 1990: Ch. 13.2).

This paper focuses on iconicity of cohesion in variation of lexical and analytic causatives in English.[1] Consider (3a), which contains an analytic (or periphrastic, in some works) causative *cause* + *to* V, and (3b), which illustrates the category of lexical causatives with a transitive example of the verb *melt*.

(3) a. Greenhouse gas emissions cause the ice caps to melt.

b. The heat from the lamp melts the ice.

Both causative constructions designate a causative situation that consists of a causing and caused events. Analytic causatives represent the events separately. The causing event is represented in a very abstract way by the first verb (*caused*), which is called the causal predicate, and the caused event is represented by the second verb, or the effected predicate (*to melt*). Other examples of analytic causatives are *make* + V, *have* + V and *get* + *to* V. In lexical causatives, the causing and caused events are merged in one predicate, as in $melt_{TR}$. Other examples of lexical causatives in English are verbs $break_{TR}$, *cut*, $bend_{TR}$, *send* and *give*. In this paper, we will use the term Causer to refer to the entity that brings about the causing event, and the term Causee to designate the entity that brings about the caused event. A more detailed definition of analytic and lexical causatives as comparative concepts is provided in Section 3.

The structural difference between analytic and lexical causatives is usually interpreted iconically (e.g. Haiman 1983). The formal merge of the causing and caused events in lexical causatives also means that these events are more integrated conceptually than when they are expressed by two different words in analytic causatives. In (3a), the use of an analytic causative

[1] Morphological causatives (e.g. Turkish *öl-dür-* "to kill" from *öl-* "to die".) are semantically and syntactically intermediate with regard to analytic and lexical causatives (Comrie 1981: Ch. 8). Since morphological causatives are not productive in English, they are left out of the discussion.

suggests that the greenhouse gas emissions trigger some processes in the atmosphere that cause global warming, which in its turn results in the melting of ice caps. Thus, the causation is less direct (and therefore less evident, which makes it possible to deny climate change) than in Example (3b), which designates more direct causation.

There is also semantic variation among English analytic causatives. For example, one can expect a higher degree of integration of the causing and caused events encoded by the English causatives *make* + V and *have* + V, and a lower degree of conceptual integration in *cause* + *to* V and *get* + *to* V because the former contain a bare verb, and the latter a *to*-infinitive, which increases the formal distance between the causing and caused events (Givón 1990: 974; Fischer 1995; Hollmann 2004). At the same time, analytic causatives are believed to convey a higher degree of conceptual integration than such constructions as in (4), where the causing and caused events are represented by different finite clauses. The causing events are in italics, and the caused events are underlined.

(4) I *made sure* <u>that Jay-Z was helping Beyonce out</u>', Obama reveals how *he ensures* <u>the first-time dad is pulling his weight with the new baby</u>.[2]

With animate Causees, the semantic difference between more compact and less compact causatives is most commonly related to the degree of agentivity, volitionality or control on the part of the Causee. Consider (5):

[2] http://www.dailymail.co.uk/news/article-2220850/I-sure-Jay-Z-helping-Beyonce-Obama-reveals-ensures-time-dad-pulling-weight-new-baby.html (last accessed 29.01.2015)

(5) a. He made the children lie down.

b. He laid the children down.

Following Haiman (1983: 784), (5a) is possible if the children are awake and respond to the Causer's command by performing the action themselves. They are agentive participants. In contrast, (5b) is appropriate when the children are asleep and unconscious or unwilling to comply, so that the Causer is the main source of energy in bringing about the result, whereas the Causee is non-agentive.

With inanimate Causees, the difference often lies in the degree of spatiotemporal integration of the causing and caused events. Consider Example (6):

(6) Haiman 1983: 784

a. I caused the cup to rise to my lips.

b. I raised the cup to my lips.

In this case, (6a), unlike (6b), suggests an absence of physical contact between the Causer (the speaker) and the Causee (the cup) – for example, by telekinesis. This contrast is possible when the Causee is animate, as well. Consider a famous example from Fodor (1970) in (7). Due to a tighter spatiotemporal integration of the events in case of lexical causatives and a looser integration in case of analytic causatives it is possible to say (7a), but impossible to say (7b):

(7) Fodor 1970: 433

a. John caused Bill to die on Sunday by stabbing him of Saturday.

b. *John killed Bill on Sunday by stabbing him of Saturday.

Although these ideas have been influential in functionalist linguistics since the 1980s, the predictions made by this theory have rarely been tested quantitatively, to the best of our knowledge. Moreover, discussions of iconicity of cohesion in causatives are based on invented examples. The present study aims to fill in this gap and tests if the correlation between formal and conceptual integration can be detected in a large corpus.

3. Analytic and lexical causatives in the British National Corpus: description of the comparative concepts and the sample

3.1. Definition of analytic and lexical causatives

This case study tests if one can predict the ratios of analytic to lexical causatives in a large corpus based on semantic features that are related to the degree of conceptual integration of the causing and caused event. Since we hope that the iconicity hypothesis will be tested on different languages in the future, we use a definition of analytic and lexical causatives as concepts designed for cross-linguistic comparison (Haspelmath 2010) from Levshina (2015a):

(8)

FUNCTION: An ANALYTIC CAUSATIVE designates a causative event, which involves a causing event (or state) and a caused event (or state), and their participants, most importantly, the Causer and the Causee. The Causer initiates or is responsible for the causing event, whereas the Causee is the entity that brings about the caused event. There can also be other participants involved (e.g. the Affectee, the final affected entity, cf. Kemmer & Verhagen 1994).

FORM: An ANALYTIC CAUSATIVE consists of two VERBS and their arguments. The first VERB (V1) represents the causing event and describes it in an abstract way, whereas the other VERB (V2) represents the caused event. The order of the predicates may vary. At least one argument of V2 is grammatically dependent on V1.[3]

Examples of English causatives that meet these criteria are *make X do Y*, *cause X to do Y*, *have X do Y* and *get X to do Y*. Consider examples (9) to (12).

(9) And I I er anyway it did make me ill, it *made* me *bleed*. (BNC FXX)

(10) At the same time they intensify their economic attack and *cause* Cuba *to fall* into economic difficulties. (BNC G1R)

(11) Individuals will interact in much more unpredictable and complex ways than the classical writers would *have* us *believe*. (BNC GVN)

(12) It would have been impossible to *get* her *to eat* if there was the slightest bit of tension in her. (BNC CHE)

[3] The words in small caps are comparative concepts. See Levshina (2015a), which suggests a definition of the comparative concept VERB.

Examples of causatives that are similar but do not meet all the criteria are *make sure (that) P* and *ensure (that) P*, as in (4). Another example is the so-called *into*-causative *V X into Ving* (see Stefanowitch & Gries 2003 and later works), in (13). The reason it does not meet the definition is that V1 (*provoked*) is too specific.

(13) Luckily he had photocopied it before he *provoked* her *into doing* this. (BNC AC3)

We also do not take into account *let* + V, although formally this construction meets all criteria. The reason is that including this construction would introduce an additional semantic dimension, namely, permissive vs. factitive causation (Nedjalkov 1976: Ch. 3), which might interfere with the results. As our cross-linguistic studies demonstrate, this is an important dimension of variation of causatives in many European languages (Levshina, 2016). Other constructions that are excluded are *have/get* + Past Participle, which express curative causation (from Latin *curare* 'to arrange, command'), such as *have one's hair cut*. Note that the passive uses of both analytic and lexical causatives were not taken into account, either.

Lexical causatives are easier to define. Functionally, they denote bringing about a change in state, location, configuration of another entity or its possession status. Formally, they consist of one verb, which represents both the causing event and the caused event. Examples are *break*$_{TR}$, *boil*$_{TR}$, *raise*, *give* and *send*. Lexical causatives need not consist of one typographic word only. Phrasal verbs like *break*$_{TR}$ *off* or *give away* are regarded as lexical causatives, as well. It is only essential that both events are expressed through a single predicate. In English, all lexical causatives are transitive or ditransitive verbs, but not all transitive and

ditransitive verbs are also causatives. For example, while both *break* and *hit* can be transitive, only *break* can be transitive AND causative. The verb *hit*$_{TR}$ is not causative because the semantics of hitting does not involve a change in the object of hitting.

3.2. The sample

For the present study, we created a list of 62 pairs of analytic and lexical causatives with similar or at least overlapping meanings. An example of such a pair is *break*$_{TR}$ and *cause/make/have/get* + (*to*) *break*$_{INTR}$. Consider examples in (14), which illustrate the pair CAUSE + boil$_{INTR}$/ boil$_{TR}$:

(14) a. For a brief moment, the terrible Red-Hot Smoke-Belching Gruncher *made* the lake *boil* and smoke like a volcano, then the fire went out and the awesome beast disappeared under the waves. (BNC CH9)

b. Do not *boil* the soup, especially if you add soured cream or natural yogurt. (BNC ABB)

A full list of pairs is presented in the Appendix. Although there is variation among English analytic causatives regarding the difference in the form of the infinitive (with or without *to*) and other features (e.g. Gilquin 2010; also see Section 2), we will treat them as one abstract construction, e.g. CAUSE + *break*$_{INTR}$. The pairs were identified partly on the basis of the list of causative and inchoative verbs in Haspelmath (1993), and a few more pairs were added. Most pairs contain a labile verb in English, e.g. *boil*$_{INTR}$/ *boil*$_{TR}$, but some are suppletive (e.g.

vs. *kill*/CAUSE + *die*) or contain originally morphological causatives (e.g. *raise*/CAUSE + *rise*). The choice of the causatives was motivated only by the availability of both lexical and analytic counterparts and availability of their examples in the corpus. This is why there is a strong bias towards labile verbs, in particular, those that designate externally caused events, in Levin & Rappoport Hovav's (1995) terms.

We used the British National Corpus (XML edition), syntactically parsed with the help of the Stanford Parser.[4] This corpus contains about 100 million words, representing British English of various genres, mostly written texts (90%), although there is also a relatively small spoken component (10%). All instances of the lexical causatives were identified in the BNC by searching for all forms of the verbs in the list that have a direct object. The instances of the analytic causatives were extracted by searching for the verbs *make*, *have*, *cause* and *get* followed by the non-causal verbs from the list, e.g. *rise*, *die*, *melt*$_{\text{INTR}}$. The search results for the analytic causatives were manually checked, and all spurious hits were removed. As a result, we obtained the BNC frequencies of analytic and lexical causatives for every pair. These frequencies can be found in the Appendix.

In all verb pairs, the lexical causatives were much more frequent than the analytic counterparts. The maximal ratio of analytic vs. lexical causative was observed for the pair CAUSE + *believe*/*convince* (0.097), followed by CAUSE + *swell*$_{\text{INTR}}$/*swell*$_{\text{TR}}$ (0.057) and CAUSE + *grow*$_{\text{INTR}}$/*grow*$_{\text{TR}}$ (0.046). The smallest ratios were found for the verb pairs with the labile verbs *prepare* (0.002), *improve* (0.003), *extend* (0.005) and *gather* (0.005).

4. Lexical vs. analytical causatives: Do semantic classes of causative events matter?

[4] http://nlp.stanford.edu/software/lex-parser.shtml (last accessed 25.12.2014).

In this and two following sections we will test whether speakers' choices between lexical and analytic causatives relate to the degree of conceptual integration between the causing event and the caused event. We will operationalise conceptual integration with the help of different context-free variables. This will enable us to test the form-meaning correlation. In this section, we will use the concepts of direct and indirect causation from Verhagen & Kemmer (1997). These concepts represent an important dimension of conceptual integration; direct causation means tight conceptual integration, whereas indirect causation means loose conceptual integration. Direct causation is observed when there is no "intervening source 'downstream' from the initiator" (Verhagen & Kemmer 1997: 70). In contrast, indirect causation is defined as "a situation that is conceptualized in such a way that it is recognized that *some other force* besides the initiator is the *most* immediate source of energy in the effected event." (Verhagen & Kemmer 1997: 67, the authors' emphasis).

The central element in this distinction is the main source of energy that is required in order to bring about the result. According to this criterion, all causative situations can be subdivided into three types:

- the Causer is the main source of energy in bringing about the caused event. Consider example (15), where the caused event (the window being broken) happens because of the Causer's impact (e.g. hitting with a club or throwing a stone):

(15) I *broke* a window at the new police station in Reading (BNC FR5).

- there is some other source of energy that actually brings about the caused event or substantially facilitates it. This may be a physical process that happens under specific

circumstances (e.g. *melt*$_{TR}$), a physical force like gravity or inertia (e.g. *sink*$_{TR}$ and *roll*$_{TR}$), a machine (e.g. *fly*$_{TR}$ [a drone]) or an agentive Causee (e.g. *feed*). The role of the Causer is less prominent. Consider example (16). The Titanic sunk not because of the collision with the iceberg, but because this collision created a series of holes, which allowed water to flood the compartments.

(16) Yes, I can now reveal that in a previous life I was the iceberg that *sunk* the Titanic. (C87)

- the Causee is a cogniser and the caused event is mental. In such situations, it is difficult to speak about transfer of energy and therefore decide on its source. This is why events of mental causation were coded as a separate class. Consider (17), where the caused event is seeing, which does not require any obvious transfer of energy:

(17) A boy of three picked up a terrorist car bomb in the street – and innocently took it home to *show* his mum and grandma. (BNC CH6)

All causation situations in our 62 pairs were classified into these three types.[5] We found 35 situations where the Causer is the main source of energy (e.g. *break*, *close*, *fill*, *fold*, *improve*, *open*, *prepare*, *shake*, *spread* and *split*), 22 events where the Causer is not the main

[5] The classification was performed on the basis of the most basic non-figurative meaning of the lexical causatives.

source of energy (e.g. *burn*, *drop*, *feed*, *fly*, *melt*, *roll*, *sink* and *swell*), and 5 instances of causing a mental event or state (*convince*, *remind*, *show*, *teach* and *worry*).

The next task was to investigate the relationship between these types and the use of analytic and lexical causatives. Figure 1 displays a box-and-whisker plot.[6] Each box represents one of the three types. The dots are the actual data points, i.e. the pairs of lexical and analytic causatives, which belong to one of the three event classes. The position of a pair with regard to the vertical axis corresponds to its actual log-transformed ratio of analytic to lexical causatives. That is, the higher a point, the greater the odds of the corresponding analytic causative against the lexical one. Log-transformations are commonly performed in order to rescale the values so that the score distributions could be observed more clearly in a graph. Some amount of jitter has been added to the horizontal coordinates in order to avoid overplotting. The thick lines inside the boxes represent the median values of each class. These are the central scores which separate the lower 50% of scores from the upper 50%. The lower boundaries of the boxes represent the cut-off points between the lower 25% of all scores and the remaining 75%. The upper boundaries separate the lower 75% from the remaining 25%. The spread of the 'whiskers' (the thin lines) shows the range of scores in each class, from the minimal to maximal ratio.

Figure 1 suggests that the pairs with *Energy Source* = 'Other' tend to have higher ratios of analytic to lexical causatives than the pairs with *Energy Source* = 'Causer'. As for the mental causation pairs, it is difficult to make a conclusion because there are too few data points.

[6] All statistical analyses and graphics were produced with the help of R, a free statistical environment (R Core Team 2014), including the add-on packages *mgcv*, *ggplot2*, *boot* and *coin*.

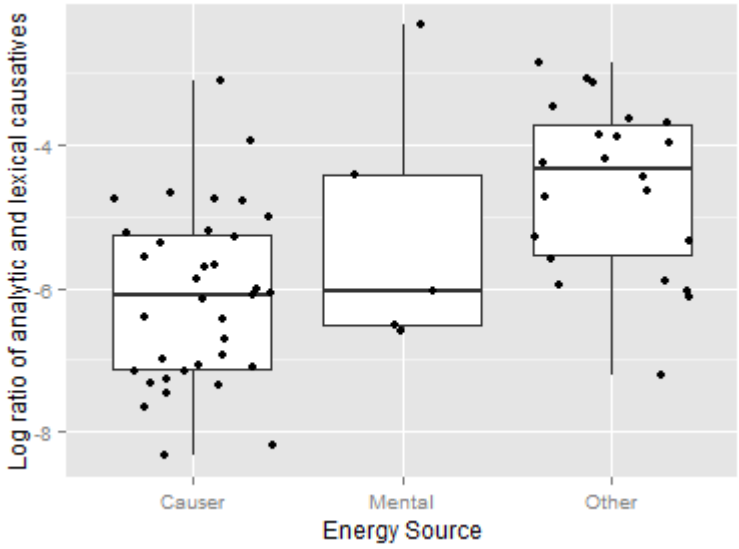

Figure 1. Box plots of types of energy source (horizontal axis) and log-transformed ratios of analytic to lexical causatives (vertical axis)

The next question is whether the differences in the ratios of analytic to lexical causatives between the three types are statistically significant. For this purpose, a logistic regression model was fitted.[7] Logistic regression is a tool for modelling a binary outcome influenced by one or more predictors. Here the outcome, or the response variable, was the frequencies of analytic and lexical causatives for each pair of causative events. The predictor was the type of causative situation. The model coefficients are shown in Table 1, along with some important general parameters that help one evaluate the quality of the model.[8]

[7] The model was quasibinomial because the dispersion turned out to be too high for a conventional binomial model

[8] Note that the pair *show*/CAUSE + *see* was removed because it had a combination of high leverage and high residual scores, and, according to the results of regression diagnostics, strongly distorted the picture.

Table 1. Estimated coefficients in a quasibinomial logistic regression model with *Energy Source* as a predictor and the odds of analytic against lexical causatives as a response.

Parametric coefficients			
Parameter	Odds ratio	Std. error	*p*-value
(Intercept)	0.002	0.27	< 0.001
Energy Source = Mental	8.00	0.46	< 0.001
Energy Source = Other	4.31	0.40	< 0.001
General model characteristics			
adjusted pseudo-R^2 = 0.13 Deviance explained = 34.4% Dispersion parameter = 28.3 (quasibinomial)			

This model shows that the differences between the causation types are indeed statistically significant and have the same direction as predicted by iconicity theory. The most important information is revealed by the estimated coefficients (under the heading 'Parametric coefficients'). The first estimated coefficient represents the so-called intercept. It shows the odds of analytic causatives compared with those of lexical causatives for the so-called reference level. The latter is the source of energy not shown in the table, *Energy Source* = 'Causer'. In general, odds greater than 1 show that the first outcome is more probable than the second outcome, whereas odds between 0 and 1 indicate that the first outcome is less frequent than the second outcome. In this model, the first outcome is an analytic causative, and the second outcome is a lexical causative. The intercept value is very small: 0.002. This makes

sense, since analytic causatives are much less frequent than lexical causatives. More exactly, the chances of analytic causatives in the situation when *Energy Source* = 'Causer' are only 0.002 times those of lexical causatives.

Let us examine the other coefficients. The coefficient of *Energy Source* = 'Other' has an estimate of 4.31. This means that the presence of another source increases the chances of analytic causatives by the factor of 4.31 in comparison with the situations when the Causer is the main energy source. This difference fully agrees with the predictions made by iconicity theory. The estimated coefficient of *Energy Source* = 'Mental' is 8.00. This means that the odds of analytic to lexical causatives are eight times greater in case of mental caused events than in cases when the Causer is the main energy source. Although we did not have any clear expectations about this contrast, it is an interesting finding that the difference is so large.

The very low *p*-values, which are displayed in the rightmost column, show that the results are statistically significant. Another important number is the adjusted pseudo-R^2. It indicates how well the model fits the data and how well it can predict the observed scores. It ranges from 0 (a completely useless model) to 1 (a perfect fit). As the relatively small score (0.13) suggests, the fit is not particularly good. This is not surprising. First, our operationalisation is based on types of situations, rather than tokens, and does not take into account possible contextual modifications. Second, there is evidence that variation of causatives in a language is multifactorial and cannot be reduced to only one semantic parameter (Levshina, 2016)..[9]

In order to be sure that the results were statistically robust, we also performed a bootstrap validation. This method is based on drawing many random samples from the data

[9] Since the number of analytic causatives was much smaller than that of lexical causatives, a quasipoisson model was also fitted. This model predicted the absolute frequency of the analytic causatives. The total frequency of both analytic and lexical causatives for each pair was used as an offset. The results led to the same conclusions.

set (the pairs of analytic and lexical causatives, in our case) with replacement. On each sample, the statistic of interest is computed. After the resamplings are done and all sample-based statistics are logged, the algorithm performs inference; for example, it can compute the 95% confidence interval for the statistics (see a more detailed explanation in Levshina 2015b: Ch. 7). The bootstrap with 9999 resamplings showed that the effects shown in the table of coefficients were significant: the 95% confidence intervals based on the resamplings did not include zero. This indicates that the differences between the types of causative situations are truly significant.

These statistical analyses demonstrate that the source of energy plays an important role in predicting the ratio of analytic to lexical causatives for a given event. Most importantly, the causation situations that involve an external source of energy have significantly higher ratios of analytic to lexical causatives than the ones where the Causer is the main source of energy. Since the presence of an external source of energy indicates a less direct causation and a lower integration of the causing and caused events, we can conclude that the iconicity-based predictions are so far borne out.

5. Corpus-driven variable: causal/non-causal alternation in English

The analysis presented in the previous section was based on a theory-informed manual classification of causation events. In contrast, this section employs a corpus-based operationalisation of event integration, namely, the ratio of causal vs. non-causal forms for each pair, e.g. *raise/rise*, *kill/die* or $break_{TR}$/$break_{INTR}$. This measure is similar to Heidinger's (2015) casualness and Haspelmath et al's (2014) (non-)causal prominence. This

operationalisation is related to the degree of spontaneity of caused events: events that are more frequently expressed by non-causal verbs are also more spontaneous than the ones that are predominantly expressed by causal verbs (Haspelmath 2005). The degree of spontaneity of the caused event is a well-known parameter in typology (Nedjalkov 1969; Haspelmath 1993).

The relationship is shown in Figure 2, where the causal/non-causal ratios are represented on the horizontal axis, and the analytic/lexical causative ratios are shown on the vertical axis. The ratios are again log-transformed for presentation purposes. The verbs are the lexical causatives, which are also the causal forms. The plot suggests that the relationship between the two variables is inverse: the greater the causal/non-causal ratio, the smaller the ratio of analytic to lexical causatives.

Let us consider the verbs located in the top left corner of Figure 2, such as *grow, explode, burst, fly, swell, bounce* and *drop*. These verbs represent events that are likely to occur spontaneously and have lower ratios of causal to non-causal uses (the horizontal axis). These events also tend to have higher ratios of analytic causatives to lexical ones (the vertical axis). In contrast, the verbs that are located in the bottom right corner, such as *improve, prepare, shake, form, hurt, increase, extend* and *open* designate events that are less likely to occur spontaneously and have higher ratios of causal to non-causal uses. These events also have lower odds of analytic to lexical causatives.

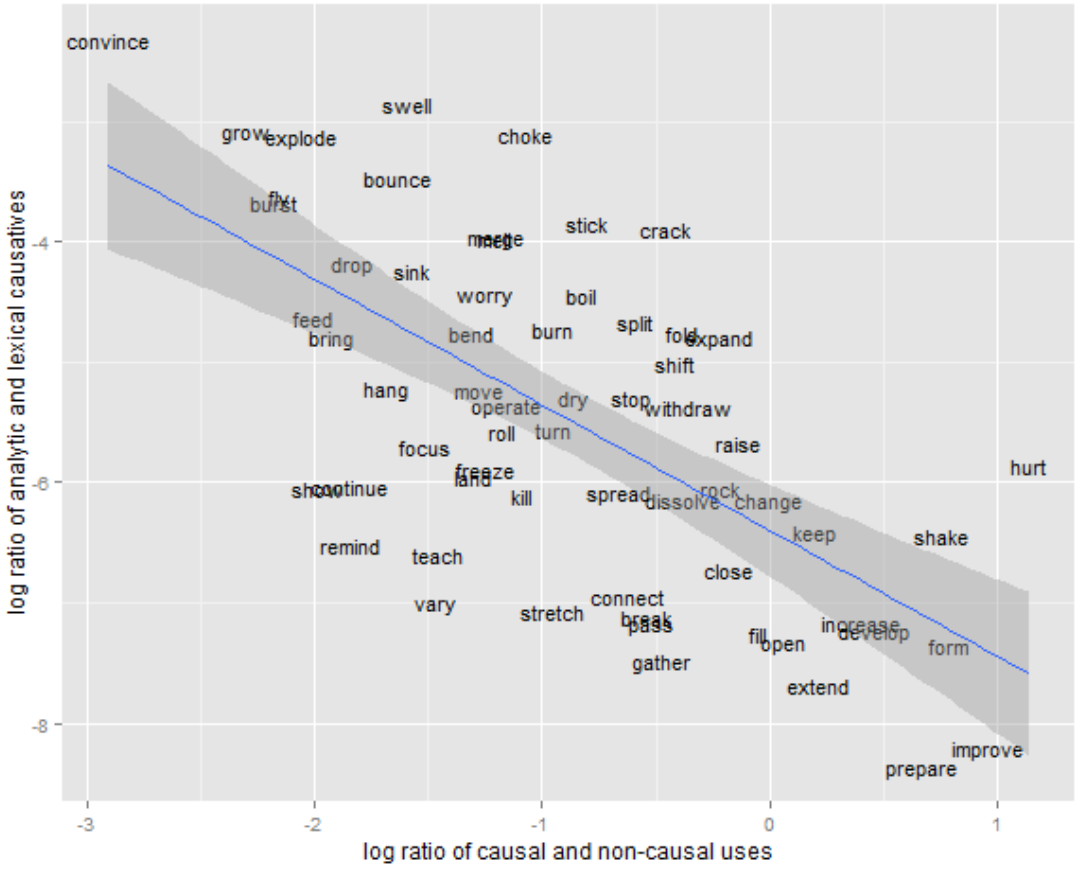

Figure 2. Negative correlation between the ratio of analytic vs. lexical causatives (vertical axis) and the ratio of causal vs. non-causal uses or verbs in the pairs (horizontal axis) in the BNC. Only the causal forms are represented. The scores are displayed on a logarithmic scale

To see whether we can rely on the results of this visual inspection, we performed a correlation test: Spearman's $\rho = -0.62$, $p < 0.001$. Spearman's ρ is the correlation coefficient, which can be from –1 (a perfect negative, or inverse correlation) to 1 (a perfect positive correlation). A zero correlation coefficient indicates the absence of any relationship between two variables. In our case, the correlation coefficient is negative ($\rho = -0.62$), which means that the relationship is inverse, as we saw in Figure 2. The very small *p*-value ($p < 0.001$) indicates that the correlation is highly significant.

The verbs located in the top left and bottom right corners, which were mentioned above, look similar to the causation types based on the source of energy, which were discussed in Section 4. In fact, the causal/non-causal ratios and the causation types are significantly correlated, as another logistic regression model shows.[10] This means that less spontaneous caused events are also associated with the Causer as the main source of energy, whereas more spontaneous events are caused by other sources. However, the correlation is not particularly strong, judging from a moderate goodness-of-fit score (pseudo-R^2 = 0.29). This means that these two operationalisations of (in)directness of causation reflect related, but different phenomena.

The results of the analyses presented in this section are in fact similar to the results obtained by Heidinger (2015) for causal alternations in French and Spanish. In his case study of analytic and lexical causatives, causation is expressed more frequently analytically when the alternating verb has a low degree of casualness, measured as the proportion of causal vs. non-causal uses of the verb.

6. Cross-linguistic evidence: inchoative-causative alternation

In this subsection, we will compare the ratios of analytic and lexical causatives in English with the results of a cross-linguistic study by Haspelmath (1993), who investigated morphological properties of 30 inchoative-causative verb pairs in 21 typologically diverse languages. Some

[10] A quasibinomial logistic model with *Causal* vs. *Non-causal* as the response variable and *Energy Source* as a predictor. The dispersion parameter was taken to be 1283.12. For *Energy Source* = 'Mental' (as compared with the reference level *Energy Source* = 'Causer'), the estimated odds ratio of *Causal* to *Non-causal* was 4.85, $p < 0.001$; for *Energy Source* = 'Other', the estimated odds ratio was 2.2, $p = 0.002$. The pseudo-R^2 was 0.29.

verbs have a basic inchoative form and a derived causative form, e.g. French *fondre* 'to $\text{melt}_{\text{INTR}}$' vs. *faire fondre* 'to melt_{TR}' and Turkish *kuru-mak* 'to dry_{INTR}' vs. *kuru-t-mak* 'to dry_{TR}'. Such verbs participate in the causative alternation. Other verbs have a basic causative form and a derived inchoative (non-causal) form and thus participate in the anticausative alternation, e.g. Russian *lomat'* 'to break_{TR}' vs. *lomat'sja* 'to $\text{break}_{\text{INTR}}$' and German *öffnen* 'to open_{TR}' vs. *sich öffnen* 'to $\text{open}_{\text{INTR}}$'.

After adding up the number of languages with one and the other alternation, Haspelmath (1993) computed the ratio of anticausative to causative alternations for every verb pair. He discovered a remarkable tendency: events that tend to occur spontaneously are usually expressed by verbs that participate more frequently in the causative alternation. Across the languages, these events are usually represented by inchoative forms that are more basic than the corresponding causative forms. Conversely, events that tend to be caused by an external agent tend to have causative forms that are more basic than the inchoative ones and therefore participate more frequently in the anticausative alternation.

We took the ratios of analytic and lexical causatives in British English and the ratios from Haspelmath (1993), excluding the pair *kill/die*, which was suppletive in the majority of languages that he considered, and pairs that had zero instances of analytic causatives in our data. This relationship is displayed in Figure 3. Again, the ratios are represented on the logarithmic scale for greater visibility. Similar to Figure 2, only the lexical causatives are displayed, which represent the corresponding pairs.

As one can see from Figure 3, the relationship between the two variables is inverse. That is, the greater the cross-linguistic anticausative/causative alternation ratio, the smaller the ratio of analytic to lexical causatives in British English, and the other way round. A correlation test shows that this negative correlation is moderately strong and statistically significant:

Spearman's ρ = –0.55, p = 0.005. The negative coefficient (ρ = –0.55) corresponds to the inverse relationship that was detected with the help of the plot.

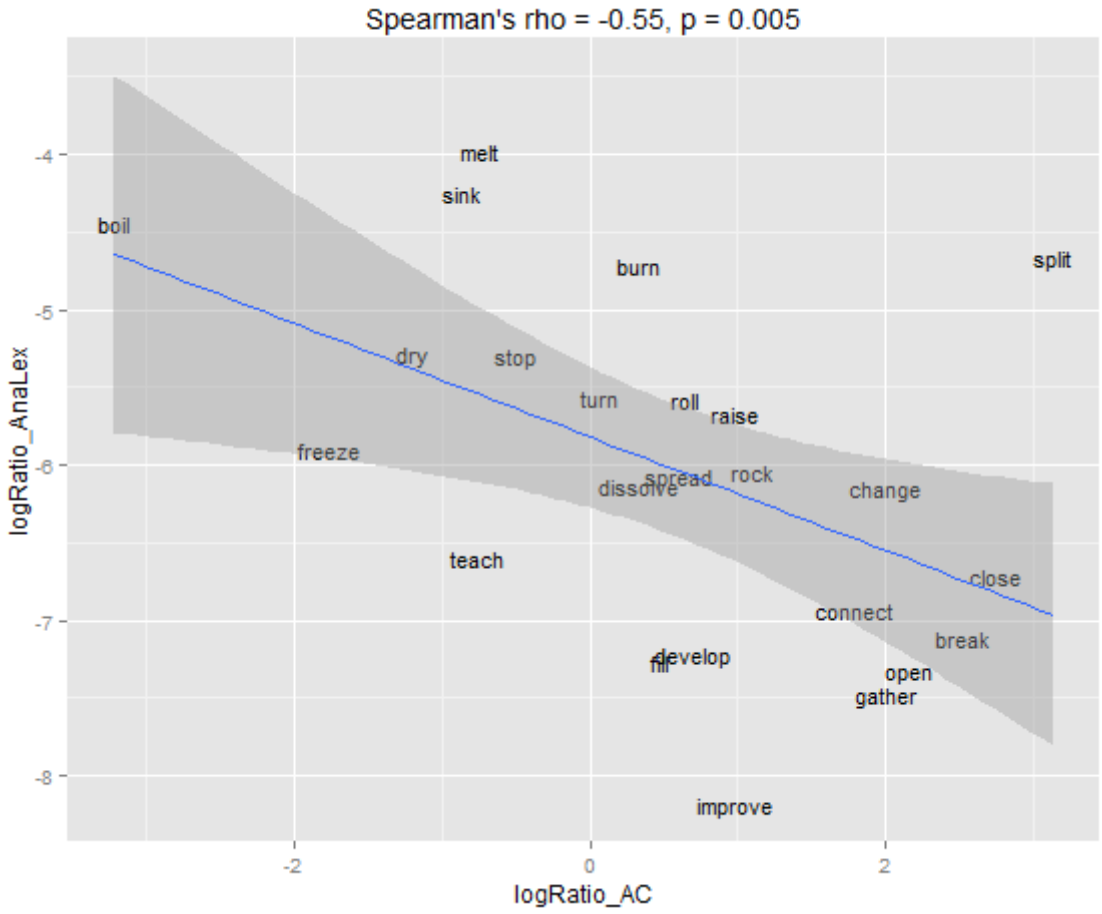

Figure 3. Negative correlation between ratios of analytic and lexical causatives in the BNC (vertical axis) and cross-linguistic anticausative/causative alternation ratios (horizontal axis) from Haspelmath (1993). The ratios are represented on a logarithmic scale

This finding can be interpreted iconically, as well. Since the caused events with a low anticausative/causative alternation ratio tend to be more spontaneous, one can expect the Causer's impact to be smaller and the causation less direct. This boosts the chances of English analytic causatives in comparison with their lexical counterparts. The caused events with a high anticausative/causative ratio are less spontaneous. Thus, the Causer's role is greater, and causation may be perceived as more direct, which increases the odds of lexical causatives.

It is quite remarkable that the use of the English causatives can be predicted with the help of a typological parameter. This finding suggests that the cross-linguistic tendencies in the conceptualization of events are very similar, if not universal (cf. Haspelmath *et al.* 2014). A mechanism that explains how these semantic similarities manifest themselves in cross-linguistic patterns of formal variation is described in Section 7.

7. Discussion: the form-meaning correlation and how to explain it

The study has focused on iconicity of cohesion in a case study of English analytic and lexical causatives. The quantitative analyses of 62 pairs of analytic and lexical causatives in the British National Corpus have yielded the following results:

1) The presence of a source of energy other than the Causer increases the chances of analytic causatives, whereas the events with the Causer as the main source of energy have a higher probability of lexical causatives being preferred. The source of energy is relevant for the distinction between indirect and direct causation (Verhagen & Kemmer 1997): the Causer is the main source of energy in situations of direct causation, whereas indirect causation involves

another source of energy. Therefore, direct causation increases the chances of lexical causatives, while indirect causation increases the chances of analytic causatives.

2) The lower the ratio of causal to non-causal uses of the verbs that express the same caused event, the higher the chances of the corresponding analytic causatives in comparison with the lexical causatives. The ratio of causal to non-causal verbs or verb uses can be regarded as an indicator of the relative spontaneity of the event (Nedjalkov 1969; Haspelmath 2005): more spontaneous events tend to have lower ratios. Therefore, the more spontaneous an event, the higher the chances of the analytic causative being chosen. These results are very similar to Heidinger's (2015) results in his study of causal alternations in French and Spanish.

3) The higher the ratio of anticausative vs. causative alternations in a sample of diverse languages from Haspelmath (1993), the lower the chances of analytic causatives in English in comparison with their lexical counterparts. The anticausative/causative alternation ratio is a typological parameter that reflects the degree of spontaneity of events. Thus, the more spontaneous an event, the greater the chances of the corresponding analytic causative.

Table 2. A summary of results

Variable	**Higher ratio of analytic causatives**	**Lower ratio of analytic causatives**
1. Source of energy required for bringing about the caused event	Some other entity is the main source of energy (indirect causation)	The Causer is the main source of energy (direct causation)
2. Causal/non-causal verb ratio in the BNC	more non-causal, fewer causal uses (more spontaneous caused events)	more causal, fewer non-causal uses (less

		spontaneous caused events)
3. Haspelmath's (1993) cross-linguistic anticausative/causative alternation ratio	more causative, fewer anticausative alternations (more spontaneous caused events)	more anticausative, fewer causative alternations (less spontaneous caused events)

A summary of these findings is provided in Table 2. As one can conclude, all three analyses converge. The more integrated the causing and the caused events – that is, the more prominent the role of the Causer and the less spontaneous the caused event – the higher the chances of the more cohesive causative form (i.e. the lexical causative) being preferred. Conversely, a lower degree of integration between the events increases the probability of the less cohesive construction (i.e. the analytic causative). Thus, the findings of our statistical analyses can be interpreted as evidence of iconic relationships between form and function.

A fundamental question that remains is how to explain this correlation. There are several options. The relationship between form and function can be represented as shown in Figure 4. This represents the main idea of iconicity theory: iconicity is a fundamental principle that directly determines language structure.

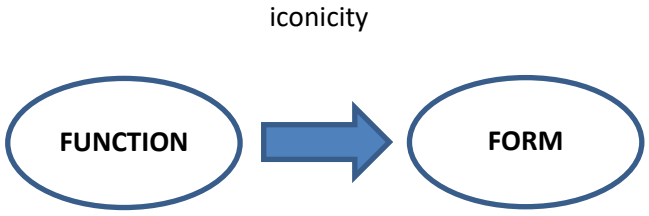

Figure 4. Iconicity as a force that determines language structure directly

However, correlation does not imply causation. An alternative explanation of the formal asymmetries has been proposed recently by Haspelmath (2008a, 2008b; Haspelmath *et al.* 2014 and other works), who performed a systematic critical re-evaluation of many classical examples of iconicity across diverse languages, including a small case study of variation of analytic and lexical causatives in English (Haspelmath 2008b: Section 6.2). The higher degree of formal cohesiveness in lexical causatives can be explained by their higher relative frequency, rather than by their semantics (Haspelmath 2008b: 22–23). This perspective builds on the works of such linguists as Zipf (1935) and Haiman (1983). The main idea can be summarised by the famous quote from Du Bois: "[g]rammars code best what speakers do most" (Du Bois 1985: 363), which means that the most economical coding mechanisms are provided by the grammars for those speech functions which speakers most often need to perform (Du Bois 1985: 362–363). This tendency can explained by the Principle of Least Effort (Zipf 1949), or the principle of economy (Haiman 1983).

The importance of frequency has been emphasised in usage-based linguistics. A variety of frequency effects have been described in the literature (see Diessel 2007 for an overview). In particular, it has been shown that high token frequency can have a conserving effect (e.g. preserving frequent irregular forms) and increase the autonomy of an expression from the related items in the paradigm or lexical class (Bybee 1985b). However, the most important frequency effect in the context of formal asymmetries is formal reduction and fusion, which can be explained by emerging neuromotor routines (e.g. Bybee 2007: 11). Some researchers have suggested that the reducing effect of frequency may be indirect and mediated by predictability or familiarity (Haiman 1983: 802). The speaker can choose a shorter expression when s/he knows that the hearer will have sufficient contextual knowledge to understand the expression. Note that this effect is highly context-dependent and relies on conditional probabilities, rather than on absolute frequencies of an expression. For example, a cross-linguistic study performed by Piantadosi *et al.* (2011) showed that the

length of a word is inversely correlated with its average conditional probability, and that this correlation is stronger than the one between word length and frequency., although the conditional probability and frequency are significantly positively correlated, too. A causal model from Haspelmath *et al.* (2014: Figure 3) is shown in Figure 5.[11]

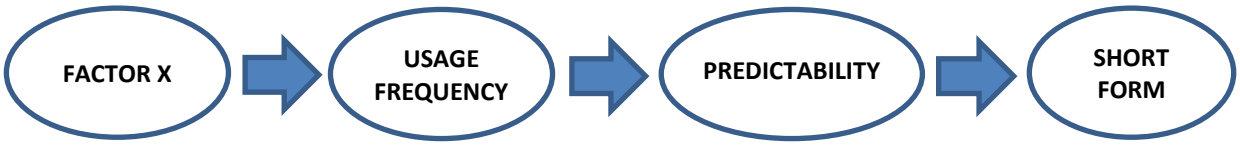

Figure 5. Frequency as a driving force of formal variation, according to a frequentist economy-based account (Haspelmath *et al.* 2014: 615).

However, the above-mentioned frequency-based explanations are not directly applicable in the case of near-synonymous causatives. It is not clear how the differences between the analytic and lexical causatives can be explained by referring to the mechanism of formal reduction that is triggered by predictability. Still, it seems that the principle of economy is at work here, too. In order to see that, one has to consider the fundamental pragmatic principles, such as Horn's (1984) Q and R Principles ('Say as much as you can' and 'Say no more than you must', respectively). These principles, in their turn, are based on Zipf's (1949: 19–23) Principle of Least Effort, which manifests itself in two opposing forces

[11] [11] According to Haspelmath *et al.* (2014), there exists a possibility that predictability depends not on frequency, but is correlated with it, while both depend on some third factor. However, in the usage-based framework, the correlation between frequency and formal length exists because frequently occurring forms are more predictable (e.g. Diessel 2007). Even if there other factors that influence the degree of predictability, we believe that the link between frequency and predictability still exists.

of the speaker's and auditor's economy. From the Q and R Principles follows that more complex and/or prolix linguistic expressions receive less marked (i.e. probable, likely, salient, etc.) interpretations via Q-based inference, and less complex expressions are associated with more stereotypical interpretations via R-based inference. An example is provided in (18):

(18) Horn 1984: 27

a. Lee stopped the car.

b. Lee got/made the car (to) stop.

The choice of the lexical causative in (18a) R-implicates that the effect was brought about in a usual way (most likely, stepping on the brake pedal), whereas the use of an analytic causative in (18b) Q-implicates that the car was stopped in some untypical way, e.g. telekinesis or pulling out the emergency brake.

Such form-meaning correspondences may become conventionalized, which results in equilibrium, known as the division of pragmatic labour (Horn 1984: 22–23). In this case, an efficient form-meaning pairing is achieved, so that the less complex constructional schema is associated with more typical meanings, and the more complex one with more marginal ones.[12] Such pragmatic principles based on the principle of economy act as an 'invisible hand' in language evolution through innumerable instances of language use, similar to the process of natural selection (cf. Keller 1994).[13]

[12] In his critique of Haspelmath's (2008) analysis of inalienable vs. alienable possession constructions in terms of economy, Croft claims that the semantic differences between the constructions can only be explained by iconicity (Croft 2008: 54). However, we believe that the economy account can be regarded as valid if one considers the pragmatic mechanisms outlined above.

[13] See also Bergen *et al*. (2012), who show in an experiment that people choose more costly (literally) utterances for less frequent meanings, whereas less costly utterances are selected for more frequent meanings. Notably, this happens in the absence of prior linguistic conventions.

An illustration of this principle in action can be found in Haspelmath *et al.* (2014: 595). The French *fondre*, which originated from Latin *fundere* 'pour', developed the meaning 'melt$_{TR}$', e.g. to melt iron. The non-causal meaning 'melt$_{INTR}$' was expressed by the reflexive form *se fondre*. With time, this meaning was extended to other kinds of melting, most importantly, melting of ice. Since we speak more often about non-causal, spontaneous melting than about causal melting, *fondre* without the reflexive *se* became increasingly used intransitively, and the causal event (i.e. melt something) is commonly expressed by *faire fondre*. Although Haspelmath et al. (2014) do not mention the pragmatic principles in their explanation, this example provides an illustration of how form and meaning can be re-mapped when the pragmatic mechanisms come into play.

We believe that the case of analytic and lexical causatives is an instance of economy effects based on pragmatic division of labour. In the case of the English causatives, we see a 100% match between form and frequency: all lexical causatives in our sample are much more frequent than their analytic counterparts. Moreover, in a random sample of 20,000 tokens from twenty BNC texts of diverse genres, one finds that the ratio of analytic to lexical causatives is approximately 1 to 31 (with the actual numbers 11 and 340, respectively). This makes frequency a perfect predictor of the formal asymmetry.

The frequency-based account of the variation of causatives can be further supported by typological evidence. Following the line of argumentation in Haspelmath (2008b: 23–24), we inspected our data base of causatives in 83 typologically diverse languages that are spoken in different parts of the world, and found 25 instances of contrasting causatives expressing different degrees of integration of the causing and the caused event, which correspond to direct vs. indirect, contact vs. distant, factitive vs. permissive or assistive causation, etc. Remarkably, the majority of these contrasting constructions (14 occurrences) differ in length, rather than cohesiveness (cf. Haspelmath 2008b: 23). Consider (19), an example from Hindi:

(19) Hindi (Kulikov 1993: 130)

paṛh-nā ‘to study’

paṛh-ā-nā ‘to teach’ (contact causation)

paṛh-vā-nā ‘to have [someone] to study’ (distant causation)

In nine cases, length correlates with cohesion.[14] The longer forms are also less cohesive, and vice versa. Consider (20) and (21). The element designating the causing event is in bold, and the element representing the caused event is underlined:

(20) Korean (Patterson 1974: 9–10)

a. *emeni-ka* *Yenghi-eykey* *say-os-lul*

mother-SUBJ Yenghi-IO new-clothes-DOBJ

*<u>ip</u>-**I**-ess-ta*.

wear-CAUS-PAST-DEC

‘Mother caused Yenghi to wear the new clothes.’

b. *emeni-ka* *Yenghi-eykey* *say-os-lul*

mother-SUBJ Yenghi-IO new-clothes-DOBJ

[14] Unfortunately, the information available in reference grammars and theoretical literature seldom allows one to compare the levels of productivity of the shorter and longer forms, which can be regarded as another manifestation of cohesion. In those cases when this information is available, the longer causative morphemes tend to be more freely distributed across different verbs classes than the shorter ones (cf. Tariana in Aikhenvald 2000: 161, Amharic in Amberber 2000: 320–321, Dulong and Rawang in LaPolla 2000: 297–299).

<u>ip</u>-key ***ha**-ess-ta*.

wear-COMP CAUS-PAST-DEC

'Mother caused Yenghi to wear the new clothes.'

(21) Mixtec (Hinton 1982: 356)

a. ***s**-<u>kee</u>*

CAUS-eat

'Feed him (by putting food directly into the mouth).'

b. ***sáʔà*** *hà* *nà* *<u>kee</u>*

CAUSE NOM OPT eat

'Make him eat (i.e. prepare things so he may eat).'

In such cases, the length of causative morphemes and their cohesiveness go together. This is perfectly natural: short expressions tend to be bound because they cannot stand on their own (Haspelmath 2008b: 18). In two remaining cases, no difference in either length or cohesion was detected.

Thus, according to our data, the longer causative forms are either less cohesive or, in the majority of cases, just as cohesive as the shorter forms. However, the less cohesive forms are *always* longer than the more cohesive forms. From this follows that length is more relevant for variation of causatives than the degree of cohesion and that the frequency model should be preferred to the iconicity explanation.

However, this purely economy-based explanation leaves two important questions open. The first one is how to explain the systematic cross-linguistic isomorphism between form and function supported by the statistical analyses presented in Sections 4–6. The second question is what motivates frequency asymmetries, i.e. what is Factor X in Figure 5. An obvious suggestion is to consider the frequency of objects or events in the world, or referential frequency. Some support for this idea comes, quite unexpectedly, from Chomsky. He is reported to use the following argument against probabilistic and corpus-based models of grammar:

> "As Chomsky himself stated so amusingly, the sentence *I live in New York* is fundamentally more likely than *I live in Dayton, Ohio* purely by virtue of the fact that there are more people likely to say the former than the latter." (McEnery & Wilson 2001: 10)

A corpus analysis performed by Stefanowitsch (2005) shows that the proportional frequencies of the sentences *I live in New York* and *I live in Dayton, Ohio* (in different modifications, e.g. *I live in New York/New York City/NYC*, etc.) very closely match the proportional sizes of the population in the two cities. This example shows that extralinguistic reality may indeed determine frequency asymmetries of linguistic constructions.

There is a danger, however, in trying to explain all linguistic frequencies by the frequencies of their referents. For example, Taylor (2012: 150–151) observes asymmetries that cannot be explained referentially. He finds that the sentence *He lives in New York* is twice as frequent as the sentence *She lives in New York* in Google, although one may expect these sentences to have approximately equal frequencies. Another example is the pair *He lives in New York* and *He lived in New York*. Although one might suppose that the people who formerly lived in New York outnumber the people who live there at present, Taylor's corpus counts show that the sentence *He lives in New York* is 1.5 times more frequent than *He lived in New*

York. One might explain these discrepancies by speakers' referential intentions: people may speak more often about males than about females due to certain cultural biases and may be more interested in a person's present place of residence than in their past places of residence. From this follows that the frequency of referential intentions is more adequate than the simple referential frequency.

Thus, if a function is referentially prominent, the construction that expresses it will be more frequent. Can one find any support for this explanation of frequency asymmetries? According to Taylor, there is no way this claim can be proven other than by looking at the linguistic frequencies: "The argument in terms of speakers' referential intentions turns out to be irredeemably circular" (Taylor 2012: 151). Still, there seems to be some indirect evidence. First, direct physical causation is very similar to force-related image schemata in Cognitive Linguistics, which represent pre-conceptual and pre-linguistic Gestalts, such as the image schema of compulsion (Johnson 1987: 45). This may be interpreted as indirect evidence of experiential basicness of direct causation. Second, one cannot ignore the fact that languages display striking similarities in encoding direct causation. All languages seem to have the transitive construction, which prototypically encodes a volitional Agent, who changes the state of a non-volitional Patient (Hopper & Thompson 1980; Næss 2007). As for less direct causation, there is substantial variation, both between and within languages, which employ causative suffixes, prefixes or auxiliaries and diverse kinds of complementation patterns. This may be interpreted as another indicator that direct causation belongs to the core of human experience.

Thus, if function determines frequency and predictability, which, in its turn, determines the form-meaning mapping according to the principle of economy, one can conclude that iconicity of cohesion (cf. Figure 4) is epiphenomenal in the sense that it does not have a direct influence on the linguistic form.

8. Concluding remarks and outlook

This paper has tested the predictions of iconicity theory based on different types of empirical evidence from English lexical and analytic causatives. The statistical analyses demonstrate that there is indeed a correlation between form and function: more cohesive forms express more direct causation, and less cohesive forms express less direct causation. One might argue, of course, that it is misleading to test one predictor only, since there may be other variables, which confound the effects found in this study. Indeed, another quantitative corpus-based analysis supports the idea that the variation of causatives in European languages is multifactorial (Levshina 2016). However, the results of that study also suggest that the variables related to the integration of events play a very important role in explaining the use of analytic and lexical causatives in most European languages, including English.

This paper has also suggested a causal mechanism that may explain this correlation. This mechanism involves an indirect causal link between function and form, via usage frequency and predictability, and is based on the principle of economy. Although testing and elaboration of this hypothesis is a task for future investigation, some supporting evidence is already available. For example, an experimental study of artificial language learning (Levshina, In prep.) demonstrates that artificial language learners tend to prefer shorter causative forms when describing more frequent causal events, in the settings where no iconic motivation is involved. Another preliminary investigation by the author shows that typologically diverse languages encode more frequent types of causation by more compact forms, and less frequent ones by less compact forms. Importantly, not all causation types can be reduced to the distinction between direct and indirect causation, which means that the principle of economy has a larger predictive power than the principle of iconicity.

Appendix: Verb pairs

Lexical (LC)	Analytic (AC)	Frequency LC	Frequency AC
$bend_{TR}$	CAUSE + $bend_{INTR}$	698	6
$boil_{TR}$	CAUSE + $boil_{INTR}$	339	4
$bounce_{TR}$	CAUSE + $bounce_{INTR}$	193	6
$break_{TR}$	CAUSE + $break_{INTR}$	6138	5
$burn_{TR}$	CAUSE + $burn_{INTR}$	1234	11
$burst_{TR}$	CAUSE + $burst_{INTR}$	238	6
$change_{TR}$	CAUSE + $change_{INTR}$	12159	26
$choke_{TR}$	CAUSE + $choke_{INTR}$	201	9
$close_{TR}$	CAUSE + $close_{INTR}$	4949	6
$connect_{TR}$	CAUSE + $connect_{INTR}$	1031	1
$continue_{TR}$	CAUSE + $continue_{INTR}$	3722	9
convince	CAUSE + believe	1839	179
$crack_{TR}$	CAUSE + $crack_{INTR}$	539	11
$develop_{TR}$	CAUSE + $develop_{INTR}$	12335	9
$dissolve_{TR}$	CAUSE + $dissolve_{INTR}$	463	1
$drop_{TR}$	CAUSE + fall	4183	64
dry_{TR}	CAUSE + dry_{INTR}	790	4
$expand_{TR}$	CAUSE + $expand_{INTR}$	1685	14

explode$_{TR}$	CAUSE + explode$_{INTR}$	181	8
extend$_{TR}$	CAUSE + extend$_{INTR}$	4307	2
feed	CAUSE + eat	1852	18
fill$_{TR}$	CAUSE + fill$_{INTR}$	4282	3
fly$_{TR}$	CAUSE + fly$_{INTR}$	862	23
focus$_{TR}$	CAUSE + focus$_{INTR}$	892	3
fold$_{TR}$	CAUSE + fold$_{INTR}$	583	5
form$_{TR}$	CAUSE + form$_{INTR}$	10846	7
freeze$_{TR}$	CAUSE + freeze$_{INTR}$	364	1
gather$_{TR}$	CAUSE + gather$_{INTR}$	1754	1
grow$_{TR}$	CAUSE + grow$_{INTR}$	1611	74
hang$_{TR}$	CAUSE + hang$_{INTR}$	1287	7
hurt$_{TR}$	CAUSE + ache	1758	5
improve$_{TR}$	CAUSE + improve$_{INTR}$	7221	2
increase$_{TR}$	CAUSE + increase$_{INTR}$	10362	8
keep	CAUSE + stay	22449	37
kill	CAUSE + die	7205	16
land$_{TR}$	CAUSE + land$_{INTR}$	768	2
melt$_{TR}$	CAUSE + melt$_{INTR}$	321	6
merge$_{TR}$	CAUSE + merge$_{INTR}$	260	5
move$_{TR}$	CAUSE + move$_{INTR}$	7787	42
open$_{TR}$	CAUSE + open$_{INTR}$	10568	7
operate$_{TR}$	CAUSE + operate$_{INTR}$	2325	11
pass$_{TR}$	CAUSE + pass$_{INTR}$	6495	5

prepare$_{TR}$	CAUSE + prepare$_{INTR}$	4228	1
raise	CAUSE + rise	13038	45
remind	CAUSE + remember	4056	6
rock$_{TR}$	CAUSE + rock$_{INTR}$	420	1
roll$_{TR}$	CAUSE + roll$_{INTR}$	1063	4
shake$_{TR}$	CAUSE + shake$_{INTR}$	5616	9
shift$_{TR}$	CAUSE + shift$_{INTR}$	1355	9
show	CAUSE + see	25434	60
sink$_{TR}$	CAUSE + sink$_{INTR}$	488	7
split$_{TR}$	CAUSE + split$_{INTR}$	851	8
spread$_{TR}$	CAUSE + spread$_{INTR}$	1737	4
stick$_{TR}$	CAUSE + stick$_{INTR}$	1463	31
stop$_{TR}$	CAUSE + stop$_{INTR}$	8188	41
stretch$_{TR}$	CAUSE + stretch$_{INTR}$	1169	1
swell$_{TR}$	CAUSE + swell$_{INTR}$	157	9
teach	CAUSE + learn	4398	6
turn$_{TR}$	CAUSE + turn$_{INTR}$	11710	45
vary$_{TR}$	CAUSE + vary$_{INTR}$	1080	1
withdraw$_{TR}$	CAUSE + withdraw$_{INTR}$	1506	7
worry$_{TR}$	CAUSE + worry$_{INTR}$	924	11

Corpus

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

References

Aikhenvald, Alexandra Y. 2000. Transitivity in Tariana. In R.M.W. Dixon & Alexandra Y. Aikhenvald (eds.), 145–172.

Amberber, Menigstu. 2000. Valency-changing and valency-encoding devices in Amharic. In R.M.W. Dixon & Alexandra Y. Aikhenvald (eds.), 312–332.

Bergen, Leon, Noah D. Goodman & Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In Naomi Miyake, David Peebles & Richard P. Cooper (eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 120–125.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: a multivariate analysis. In Isabelle Bril (ed.), *Clause-hierarchy and clause-linking: the syntax and pragmatics interface*, 51 - 101. Amsterdam: Benjamins.

Bybee, Joan L. 1985a. Diagrammatic iconicity in stem - inflection relations. In John Haiman (ed.), 11–47.

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

Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford: OUP.

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

Comrie, Bernard & Maria Polinsky (eds.). 1993. *Causatives and transitivity*. Amsterdam: Benjamins.

Croft, William. 2008. On iconicity of distance. *Cognitive Linguistics* 19(1). 49–57.

Diessel, Holger. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25. 108–127.

Diessel, Holger. 2008. Iconicity of sequence. A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics* 19. 457–482.

Dixon, R.M.W. & Alexandra Y. Aikhenvald (eds.). 2000. *Changing valency: Case studies in transitivity*. Cambridge: CUP.

Du Bois, John. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax*. Amsterdam: Benjamins, 343–365.

Fischer, Olga. 1995. The distinction between "to" and bare infinitival complements in late Middle English. *Diachronica* XII(1). 1–30.

Fodor, Jerry. 1970. Three Reasons for Not Deriving "Kill" from "Cause to Die." *Linguistic Inquiry* 1(4). 429–438.

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

Givón, Talmy. 1980. The binding hierarchy and the typology of complements. *Studies in Language* 4(3). 333–377.

Givón, Talmy. 1990. *Syntax: A functional-typological introduction.* Vol. II. Amsterdam: John Benjamins.

Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.

Haiman, John (ed.). 1985. *Iconicity in syntax.* Amsterdam: Benjamins.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), 87–120.

Haspelmath, Martin. 2005. Universals of causative verb formation. Talk given at LSA Institute, MIT, LSA.206, 2 August 2005.

Haspelmath, Martin. 2008a. Creating economical morphosyntactic patterns in language change. In Jeff Good (ed.), *Linguistic Universals and Language Change*, 185–214. Oxford: OUP.

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

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

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

Heiddinger, Steffen. 2015. Causalness and the encoding of the causative-anticausative alternation in French and Spanish. *Journal of Linguistics* 51(3). 562-594.

Hinton, Leanne. 1982. How to cause in Mixtec. In *Proceedings of the Eighth Annual Meeting of the Berkley Linguistics Society*, 354–363.

Hollmann, Willem B. 2004. The iconicity of infinitival complementation in Present-day English causatives. In Constantino Maeder, Olga Fischer, & William J. Herlofsky (eds.), *Outside-in – inside-out. Iconicity in language and literature*, 287–306. Amsterdam: Benjamins.

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

Horn, Laurence R. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin (ed.), *Meaning, form, and use in context: Linguistic applications*, 11–42. Washington DC: Georgetown University Press.

Johnson, Mark. 1987. *The Body in the Mind: The bodily basis of meaning, imagination and reason*. Chicago: University of Chicago Press.

Keller, Rudi. 1994. *On Language Change: The Invisible Hand in Language*. London: Routledge.

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

Kulikov, Leonid. 1993. The "second causative": A typological sketch. In Bernard Comrie & Maria Polinsky (eds.), 121–154.

LaPolla, Randi J. 2000. Valency-changing derivations in Dulong/Rawang. In R.M.W. Dixon & Alexandra Y. Aikhenvald (eds.), 282–311.

Levin, Beth & Malka Rappoport Hovav. 1995. *Unaccusativity: at the syntax-semantics interface*. Cambridge, MA: MIT Press.

Levshina, Natalia. In preparation. What can artificial language learning tell us about language universals?

Levshina, Natalia. 2015a. European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2). 487–520.

Levshina, Natalia. 2015b. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: Benjamins.

Levshina, Natalia. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50(2). 507–542.

McEnery, Tony & Andrew Wilson. 2001. *Corpus Linguistics. An Introduction*. Second edition. Edinburgh: Edinburgh University Press.

Næss, Åshild. 2007. *Prototypical transitivity*. Amsterdam: John Benjamins.

Nedjalkov, Vladimir P. 1969. Nekotoryje verojatnostnyje universalii v glagol'nom slovoobrazonanii [Some probabilistic universals in verbal derivation]. In I.F. Vardul' (ed.), *Jazykovyje universalii i lingvističeskaja tipologija*, 106–114. Moscow: Nauka.

Nedjalkov, Vladimir P. 1976. *Kausativkonstruktionen*. Tübingen: TBL.

Patterson, Betty S.J. 1974. *A study of Korean causatives*. In *Hawaii Working Papers in Linguistics* 6.4, 1–52. Honolulu: Department of Linguistics, University of Hawaii.

Piantadosi, Steven T., Harry Tilly & Edward Gibson. 2011. Word lengths are optimized for efficient communication. *PNAS* 108(9). 3526–3529. Retrieved from http://www.pnas.org/content/108/9/3526.full (last accessed 21.09.2015).

R Core Team. 2014. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* Vienna. Retrieved from http://www.R-project.org/ (last accessed 25.12.2014)

Rosenbach, Anette. 2003. Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 379–411. Berlin: Mouton de Gruyter.

Stefanowitsch, Anatol. 2005. New York, Dayton (Ohio), and the Raw Frequency Fallacy. *Corpus Linguistics and Linguistic Theory* 1–2. 295–301.

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

Steger, Maria & Edgar W. Schneider. 2012. Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In Bernd Kortmann & Benedikt Szmrecsanyi (eds.), *Linguistic complexity: Second language acquisition, indinization, contact*, 156–191. Berlin: De Gruyter.

Talmy, Leonard. 2000. *Toward a cognitive semantics*. Cambridge, MA: MIT Press.

Taylor, John. 2012. *The Mental Corpus: How language is represented in the mind*. Oxford: OUP.

Verhagen, Arie & Suzanne Kemmer. 1997. Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics* 24. 61–82.

Zipf, George K. 1935. *The psycho-biology of language*. Cambridge, MA: MIT Press.

Zipf, George K. 1949. *Human behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.