Coding causal-noncausal verb alternations: a form-frequency correspondence explanation*

Martin Haspelmath, Andreea Calude, Michael Spagnol, Heiko Narrog, Elif Bamyacı

We propose, and provide corpus-based support for, a usage-based explanation for cross-linguistic trends in the coding of causal-noncausal verb pairs, such as *raise/rise*, *break* (tr.)/*break* (intr.). While English mostly uses the same verb form both for the causal and the noncausal sense (labile coding), most languages have extra coding for the causal verb (icausative coding) and/or for the noncausal verb (anticausative coding). Causative and anticausative coding is not randomly distributed (Haspelmath 1993): Some verb meanings such as 'freeze', 'dry' and 'melt' tend to be coded as causatives, while others such as 'break', 'open' and 'split' tend to be coded as anticausatives. We propose an explanation of these coding tendencies on the basis of the form-frequency correspondence principle, which is a general efficiency principle that is responsible for many grammatical asymmetries, ultimately grounded in predictability of frequently expressed meanings. In corpus data from seven languages, we find that verb pairs for which the noncausal member is more frequent tend to be coded as anticausatives, while verb pairs for which the causal member is more frequent tend to be coded as causatives. Our approach implies that linguists should not rely on form-meaning parallelism when trying to explain cross-linguistic or language-particular patterns in this domain.

1. Overview

This paper proposes a new explanation for a cross-linguistic tendency that was first discussed in Croft (1990) and first documented in Haspelmath (1993): Certain types of causal-noncausal verb pairs such as those in (1a-c) show a greater propensity for causative coding (i.e. with an extra marker on the verb), while others such as those in (2a-c) show a greater propensity for anticausative coding (i.e. with an extra marker on the noncausal verb). Intuitively, the verb pairs in (1) can be characterized as having a more "spontaneous" core-event, while the verb pairs in (2) have a less spontaneous core-event (where by *core-event* we mean the event that is shared by the meanings of both verbs).

(1) more spontaneous core-events: 'dry', 'melt', 'freeze'

	causal	noncausal
a.	'dry (tr.)'	'dry (intr.)'
b.	'melt (tr.)'	'melt (intr.)'
c.	'freeze (tr.)'	'freeze (intr.)'

(2) less spontaneous core-events: 'break', 'open', 'split'

	causal	noncausal
a.	'break (tr.)'	'break (intr.)'
b.	'open (tr.)'	'open (intr.)'
c.	'split (tr.)'	'split (intr)'

Thus, we typically find situations such as those in Japanese and Swahili, shown in (3), where the more spontaneous 'freeze' pair shows causative coding (marked by *-ase* in

* We are grateful to three anonymous reviewers for the journal of Linguistics as well as to Bernard Comrie for very useful comments on this paper. In addition, we are grateful to the audiences in several places where we presented this work: at the Max Planck Institute for Evolutionary Anthropology (which also deserves thanks for bringing several of the authors together), at the 4th UK Cognitive Linguistics Conference (London), at the Societas Linguistica Europaea 2012 in Stockholm, and at the Syntax of the World's Languages 5 in Dubrovnik.

Japanese and *-isha* in Swahili), while the less spontaneous 'break' pair shows anticausative coding (marked by *-e* in Japanese and *-ika* in Swahili).

(3)	causal	noncausal	
Japanese	*koor-**ase**-*	*koor-*	causative coding
Swahili	*gand-**isha***	*ganda*	
	'freeze (tr.)'	'freeze (intr.)'	
Japanese	*war-*	*war-**e**-*	anticausative coding
Swahili	*vunja*	*vunj-**ika***	
	'break (tr.)'	'break (intr.)'	

The basic idea that is new here is that in addition to form-meaning parallelisms, languages also exhibit FORM-FREQUENCY CORRESPONDENCES, and that these are actually more important in determining grammatical coding such as causative and anticausative coding of causal-noncausal verb pairs. In verb pairs of type (1), the noncausal member is relatively more frequent, so that the causal member tends to be coded overtly (as causative). In verb pairs of type (2), the causal member is relatively more frequent, so that the noncausal member tends to be coded overtly (as anticausative). These claims are supported by corpus data from seven diverse languages.[1]

It thus turns out that coding types in causal-noncausal alternations broadly follow the same principles of economic expression that are found widely elsewhere in grammatical structure (e.g. Greenberg 1966, Haiman 1983, Croft 2003: ch. 4, Hawkins 2004: ch. 4).

Our form-frequency correspondence account solves a problem that has vexed many researchers who assumed some kind of form-meaning parallelism principle. From such a perspective, anticausatives are puzzling because they seem to have "less meaning", but more form, than their causal counterparts (see, e.g., Haspelmath 1993: §1), or they are formally derived but semantically basic. And if one turns the analysis around and says that the noncausals are semantically derived from the causals (e.g. Levin & Rappaport Hovav 1995), then causatives are puzzling, because they are formally derived but semantically basic. On our form-frequency correspondence account, these puzzles disappear, because the trend in the formal patterns is exactly as expected: In general, grammatical coding asymmetries are sensitive to frequency asymmetries, not to semantic derivation, to "semantic markedness" or to semantic complexity (Haspelmath 2008a).

Most earlier work on causal-noncausal alternations has had rather different goals. Unlike authors such as Levin & Rappaport Hovav (1995), Piñón (2001), Doron (2003), Alexiadou et al. (2006), Schäfer (2008), Koontz-Garboden (2009) (see Schäfer 2009 for an overview), we do not aim to provide elegant ("formal") descriptions or analyses of particular languages. Our goal is restricted to explaining a cross-linguistic trend in the distribution of causatives and anticausatives, and thus our explanatory proposal is in principle compatible with all of these analytical proposals. As noted in Haspelmath (2004, 2010), the explanation of cross-linguistic trends is often orthogonal to the issue of language-particular analysis. However, to some extent many of the analytical proposals are motivated by the desire to make the presence or absence of (anti)causative coding fall out from the analysis, and this motivation disappears if we are right that the coding is best explained by frequency asymmetries and is otherwise quite possibly random (cf. Spagnol 2011: 151-152). So in this respect, our proposal can be seen as

[1] This idea was first presented in Martin Haspelmath's Linguistic Institute lectures at MIT in 2005 and was first mentioned in print in Haspelmath (2008a: 12-13). The original idea was inspired by Wright's (2001: §4.4) frequency counts for English. Samardžić & Merlo (2012), whose closely related work came to our attention after our study was completed, was inspired by Haspelmath (2008a) (as was Cysouw 2008: 388).

relevant to language-particular analysis after all, though it is beyond the scope of this paper to spell this out in detail.

In the following sections, we will lay out the basic concepts (§2), contrast the two competing explanatory principles (form-meaning parallelism and form-frequency correspondence) (§3), present the sample of 20 comparative verb meanings and the sample of seven languages (§4), and spell out our predictions (§5). Then the results of our corpus studies will be presented (§6-7), and the semantically-based approach of Levin & Rappaport Hovav will be discussed briefly (§8), before we conclude the paper (§9).

2. Basic concepts

We are interested in causal-noncausal verb pairs that exhibit some cross-linguistic variability of coding.[2] These are often called "inchoative-causative" verb pairs (following Guerssel et al. 1985, Borer 1991), but we avoid this term here, because the causal verb need not be coded as a causative, and the noncausal verb need not be semantically inchoative (i.e. it need not be a change-of-state verb containing a 'become' meaning component). We thus limit the terms *causative* and *anticausative* to their use for formal patterns,[3] while the terms *causal verb* and *noncausal verb* are used for semantic types.[4] In the context of this article, a causal verb is a verb (or verbal expression)[5] that includes a 'cause' meaning component (e.g. English *break* (tr.), which means 'cause to become broken'), while a noncausal verb is a verb that has the same basic meaning as a causal verb but lacks the 'cause' component (e.g. English *break* (intr.), which means 'become broken'). The meaning component that is shared by both verbs of a causal-noncausal pair is called the *core-event* here (this is identical to the meaning of the noncausal verb).

We focus our attention on those kinds of verb pairs that exhibit coding variability across languages (and not uncommonly within languages as well).[6] For example, the causal-noncausal pair 'wake up (tr./intr.)' can be coded in four different ways across languages:[7]

Table 1. Coding type of 'wake up' in four languages

[2] Other authors commonly talk about "alternating *verbs*" rather than *verb pairs.* This terminology is appropriate for cases like English *break* (tr.) / *break* (intr.), because its is reasonable to say that *break* is the "same" verb in both its causal and its noncausal use. But English is unusual, and most languages have two different verbs (or verb forms), so we generally speak about verb pairs.

[3] The term *anticausative* was specifically introduced for the formal pattern of overtly coded noncausals by Nedjalkov & Sil'nickij (1969) (published in English as Nedyalkov & Silnitsky 1973), and in the first 35 years was used exclusively in this way (e.g. Marantz 1984, Siewierska 1984, Comrie 1985, Haspelmath 1987, 1993, Payne 1997, Dixon & Aikhenvald 2000). More recently some linguists have used the term *anticausative* in a different, loosely semantic sense to refer (apparently) to noncausal verbs in causal-noncausal alternations (e.g. Alexiadou et al. 2006, Alexiadou 2010, Schäfer 2008, 2009, Heidinger 2010). (See Lehmann 2007 for the general problem of semantic change in grammatical terminology, with formal terms tending to acquire a loose semantic sense due to nonrigorous terminological usage.)

[4] Nichols et al. (2004: 151) use the terms *induced verb* and *plain verb* for what we call *causal verb* and *noncausal verb* respectively.

[5] In addition to verbs in the strict sense (i.e. single words), we also include multi-word verbal expressions when these are established expressions of the language, e.g. German *sich öffnen* [self open] 'open (intr.)', French *faire fondre* [make melt] 'melt (tr.)'.

[6] Note that we sometimes use the term *verb pair* in a cross-linguistic sense, to refer to 'those verb pairs in all languages that are counterparts of a given meaning'. Context should make it clear when the term is used in the language-particular sense and when it is used in the cross-linguistic sense.

[7] A fifth type, suppletion (e.g. *die/kill, learn/teach*) could be added but is quite rare. Where it occurs, we subsume it under the equipollent type (e.g. in Table 5 below).

	coding type	description of coding type	'wake up (intr.)'	'wake up (tr.)'
Lithuanian	simple/causative coding	causal is overtly marked	*pabus-ti*	*pabud-**in**-ti*
French	anticausative/simple coding	noncausal is overtly marked	***se** réveiller*	*réveiller*
German	equipollent coding	symmetrically different	*aufw**a**chen*	*aufw**e**cken*
Greek	labile (or ambitransitive) coding	causal and noncausal are identical	*ksipnó*	*ksipnó*

There are several dozen commonly represented meanings that exhibit coding variability (Haspelmath 1993), but we should not forget that most verb pairs do not exhibit any variability across languages. For example, the coding of the causal-noncausal pair 'make someone dance' / 'dance' is quite uniform: In all languages, the causal verb is overtly coded in some way (whether synthetically or analytically), whereas the noncausal verb 'dance' is simple. While logically perfectly possible, the option of coding 'make dance' as a simple root and deriving 'dance' from this as an anticausative is unattested, as far as we know. In general, when the core-event is itself agentive (i.e. when one participant is a volitional agent) and atelic, languages (virtually) never use anticausative, equipollent or labile coding.[8] This seems so natural from a form-meaning isomorphism perspective that linguists have not worried about these cases, and we will leave them aside in this paper as well. However, they are fully compatible with our account, as can be easily seen by counting the frequencies of 'make dance' and 'dance' (see also the concluding section for some discussion).

Most of the verb pairs that do exhibit variability have a core-event that denotes a change of state (i.e. that includes the component 'become'), so we could limit ourselves to causal-inchoative alternations. However, variability is also found with verb pairs denoting (potentially non-translational) motion such as '(cause to) roll', '(cause to) spin', and even sometimes with two-argument core-events such as '(cause to) learn sth'. Thus, we consider all causal-noncausal alternations in principle, but focus on those that exhibit interesting cross-linguistic differences.

3. Form-frequency correspondence vs. form-meaning parallelism

The fundamental principle of the form-frequency correspondence approach (Zipf 1935) is the principle in (4).

(4) The form-frequency correspondence principle
Languages tend to use less coding material for more frequent expressions.

This is uncontroversial for word length (e.g. Zipf 1935: 23, Bybee 2006), but it is also generally valid for grammatical patterns. The insight of the principle in (5) is originally due to Greenberg (1966) (see also Croft 2003: ch. 4, Haspelmath 2008a, 2008b).

(5) The grammatical form-frequency correspondence principle
When two grammatical patterns that differ minimally in meaning (i.e. patterns that form a

[8] Examples such as English *She walked the dog* (with agentive atelic *walk* showing labile coding) are extremely rare across languages; in fact, we are not aware of a single language other than English that has such labile verbs.

semantic opposition) occur with significantly different frequencies, the less frequent pattern tends to be overtly coded (or coded with more coding material), while the more frequent pattern tends to be zero-coded (or coded with less coding material).

This principle accounts for a wide variety of form asymmetries, for example for the fact that in the oppositions in (6) (Greenberg 1966), it is the first member that tends to be zero-coded, while the second member tends to be overtly coded.[9]

(6) singular/plural, present/future, 3rd person/2nd person, nominative/accusative, active/passive, affirmative/negative, masculine/feminine, attributive adjective/predicative adjective (including copula), positive/comparative, predicative verb/nominalized verb, action word/agent noun

Alternatively, one could approach these form asymmetries with an expectation of form-meaning parallelism, for example isomorphism of complexity, as in (7).

(7) The form-meaning complexity isomorphism principle
More complex meanings are expressed by more complex forms, i.e. by more coding material.

Such a principle has in fact often been invoked (sometimes called "iconicity of complexity"),[10] but it plainly does not work for word length (consider words such as *cat*, *car*, *child*, which are semantically more complex than their taxonomic superordinates *animal*, *vehicle* and *person*, but strongly tend to be shorter than them across languages),[11] and Haspelmath (2008a) has argued that it does not work for grammatical patterns either.[12]

Form-meaning parallelism in the sense of (7) (involving isomorphism of complexity) has been invoked only by Clark & Clark (1978: 250-51) and Haspelmath (1993), as far as we know. But Haspelmath started out with the observation that the principle cannot work in its most simplistic form: If causal verbs are semantically more complex than noncausal verbs ('cause [core-event]' vs. 'core-event'), then isomorphism predicts that anticausatives should not exist. So he argued that instead of complexity in terms of semantic decomposition, what counts is "conceptual complexity": 'freeze'-type meanings are associated with a noncausal "conceptual stereotype", while 'break'-type meanings are associated with a causal stereotype. But the status of such a "stereotype" and the way in which it causes the form asymmetries is quite unclear.

Quite a few other authors have invoked a related principle, which also relies on a close link between form and meaning:

(8) The basic-derived form-meaning parallelism principle
Derived meanings are expressed by derived forms, i.e. by forms containing

[9] Note that the difference between affixal/derivational coding (as with most plural markers) and periphrastic/syntactic coding (as with most copulas) plays no role in our account (see also n. 5). Only the overt vs. zero contrast is relevant here. Affixes and function words cannot be readily distinguished cross-linguistically anyway (Haspelmath 2011).

[10] E.g. Clark & Clark (1978: 247-251), Givón (1991: §2.2): "A larger chunk of information will be given a larger chunk of code" (see more references in Haspelmath 2008a: §3).

[11] Frank Seifart and a reviewer have pointed out to us that basic-level terms like 'cat' and 'car' may be cognitively less complex in some sense than higher-level terms (cf. Rosch 1978). We recognize this, but since this kind of "cognitive complexity" is only vaguely defined, we limit the discussion to semantic complexity in the conventional sense here. (Note also that it is quite possible that the kind of lower cognitive complexity that is associated with 'car' as opposed to 'vehicle' is actually due to frequency of use.)

[12] Alternatively, the parallelism principle in (7) has been formulated in terms of "markedness": "Marked meanings are expressed by marked forms". See Haspelmath (2006: 40) for discussion and criticism.

more coding material, or at least not less. Formal and semantic derivation cannot go in opposite directions.

Most earlier work on causal-noncausal verb alternations is concerned not with cross-linguistic trends, but with elegant language-particular description ("formal analysis"), which typically involves formulating rules that derive one verb from the other one. For example, Levin & Rappaport Hovav (1995) argue that in 'break'-type verb pairs, the causal member is basic and the noncausal member is derived, and in support of this they cite the fact that noncausal 'break' needs a special anticausative marker in many languages.[13] They thus also make a claim about languages in general (not just about English), and they presuppose that overt coding should reflect the derived nature of the meaning. Similarly, Piñón (2001) seems to presuppose that derived forms should ideally be morphologically more complex, when he says that we want an analysis that "respects the morphological facts". And Schäfer (2009: 662-63) says that theories assuming a noncausal-to-causal derivation "can easily account for" causative morphology, while they "are challenged by languages that mark (a subset of) their anticausative alternants, as these are assumed to be basic, not derived".

Thus, form-meaning parallelism seems to be widely assumed, though often implicitly. If we are right that it is primarily form-frequency correspondence that is responsible for overt coding elements in grammar, this may well mean that some of the proposed language-particular analyses will have to be adapted or rethought, at least to the extent that they rely on form-meaning parallelism.[14] But as we made clear earlier, we will not be concerned with elegant analysis of particular languages, and restrict ourselves to accounting for the cross-linguistic patterns.

Form-meaning isomorphism could plausibly be explained in terms of the general principle of iconicity, but form-frequency correspondence also has a very general explanation in terms of coding efficiency. Frequently used meanings require less expression effort than rarer meanings in any efficient information-conveying system. For example, local phone numbers are usually shorter than numbers for long-distance calls, which is efficient because local numbers are dialed more often. In the case of human language, length of coding is closely related to predictability: Hearers can afford to use shorter expressions for more predictable meanings, and more frequently expressed meanings are automatically more predictable. Thus, we can clearly identify a causal mechanism that is responsible for the principle in (5):

(9) Frequency causes predictability, which causes short form:
In human language, there are recurrent diachronic mechanisms which create patterns in which short forms are used for frequent meanings because of their predictability.

The crucial role of diachronic change was recognized by Zipf (1935) and is discussed in some detail in Haspelmath (2008b). The simplest cases are examples of abbreviations, which tend to replace full forms in proportion to the frequency of use. Abbreviations are a relatively novel (and often writing-based) mechanism, but similar processes are constantly at work throughout the language system. So when a noncausal verb develops in a language that does not fit form-frequency correspondence, there is some pressure (be it ever so slight) on speakers to modify the pattern. For example, the Latin verb *fundere* 'pour' gave rise to French

[13] "Morphological marking has a function: it is needed to indicate the nonexpression of the external source" (Levin & Rappaport Hovav 1995: 88).

[14] A reviewer suggests that common-base approaches (such as Alexiadou et al. 2006, Pylkkänen 2008, Schäfer 2008, and Spagnol 2011), which derive both verbs from a common base, are more readily compatible with the variation in coding types than approaches which assume uniform causativization or uniform decausativization. But this is only so if form-meaning parallelism is assumed. It could well be that this plays no role at all in the rules that speakers have internalized, i.e. that speakers could easily internalize rules that violate (8) (as observed by Mel'čuk (1967)).

fondre, which came to mean 'melt' at some point.[15] Originally it must have referred only to causal melting (of metal), so noncausal melting was expressed by *se fondre* (using an anticausative form). Then it was extended to other kinds of melting, specifically to melting of ice, and since in this sense we talk more about noncausal melting than about causal melting, the frequency asymmetry now contradicted form-frequency correspondence. As a result, French speakers increasingly used *fondre* (without anticausative *se*) also for noncausal melting, and for causal melting, the new causative *faire fondre* ('make melt') has become quite common. These changes were presumably introduced because they make speaking and understanding more efficient. The change has been going on over many centuries and is not complete yet, but the trend is clear: Following a drastic semantic change, there is pressure to bring the coding type in line with the usage frequency.

The relationship between coding types and frequency is thus not direct, but mediated by lengthy and complex diachronic changes which maintain the efficiency of the system. As a result, coding efficiency is only a general tendency which requires a quantitative cross-linguistic approach to be recognized.

4. Twenty causal-noncausal verb pairs and seven languages

Now in order to test our hypothesis that the coding type of causal-noncausal alternations is determined by the frequency of occurrence of the causal and the noncausal members of the pairs, we need to look at usage frequencies of a representative set of verbs in a representative set of languages. We have chosen to examine twenty verb pairs in seven languages.

The 20 verb-pair meanings are given in (10).[16]

(10) boil, freeze, dry, wake up, go out/put out (fire), sink, melt, stop, turn, burn, fill, rise/raise, improve, rock, connect, gather, open, break, close, split

The seven languages are English, Japanese, Maltese, Romanian, Russian, Swahili, and Turkish. They were partly chosen for the practical reason that we had access to corpora of these languages. These seven languages are of course not fully representative of the world's languages: Three of them are Indo-European, and five of them are spoken in Europe, so not all regions and families are represented equally. But they are quite diverse in their coding types (as shown in Table 5 below), and we think that these data suffice for an initial demonstration of our claim. We make the assumption that people tend to talk about more or less the same kinds of basic topics, and it is therefore not expected that languages differ significantly in the usage frequencies of comparable expressions. In all languages, 'head' will generally be more frequent than 'knee', 'say' will be more frequent than 'wash', and 'big' will be more frequent than 'soft', due to some aspects of human nature or the way the world is. So as long as the meanings are not culture-specific or specific to the physical environment of its speakers, a truly balanced sample of languages is not necessarily required to demonstrate universal frequency trends.

In order to test our hypothesis regarding frequency of use, we appealed to language corpora consisting of at least several million words, and wherever possible including both spoken and written linguistic samples. As always, the goal of representativeness is both an important and also partly unattainable one (how does one find a sample that represents

[15] See, e.g., the etymological information available at http:/www.cnrtl.fr/etymologie/fondre.

[16] These meanings are a subset of the 31 meanings studied in Haspelmath (1993). We reduced the set from 31 to 20 to make the task more manageable. We did not add any meanings because we want to compare Haspelmath's (1993) results with ours. (Strictly speaking, we should call these meanings "verb pair meanings", but we use the short label for convenience.)

linguistic output from every speaker, across all time and all linguistic modalities?). We hope that the surprisingly consistent results of our comparative study will be seen as confirmation for the soundness of our decisions.

Due to practical considerations, we opted for two different methods of coding the data. In some languages, distinct verb forms are used for causal and noncausal members and these could be identified by simple but exhaustive searches of entire corpora. In other languages, the same verb forms can be used in both causal and noncausal constructions and the only way to disentangle the two is through careful manual checking. Hence for three languages, namely Japanese, Russian and Swahili, it was possible to look at all the occurrences of the causal and noncausal members of the pairs in the entire corpus and conduct exhaustive searches. However, for the remaining four languages, English, Maltese, Romanian and Turkish, causal and noncausal members of the pairs had to be extracted manually, so we limited our counts to the first 50 instances of each verb pair, assuming that these verb sets would be representative (see more details about this decision at the end of §6).

The sources of our corpora are given in the Appendix, as are the results for each verb in each of the seven languages. Table 2 lists the languages included, type of data sampled, total word counts, and how the causal and noncausal verbs were identified (either exhaustively, by automatic search for all forms, or manually, by analyzing the first relevant 50 examples).

Table 2: Languages and corpus data

	LANGUAGE	DATA TYPE	MODALITY	TOTAL NUMBER OF WORDS	IDENTIFICATION
1	English	various	spoken & written	100 million	manual
2	Japanese	various	written	66 million	exhaustive
3	Maltese	various	written	100 million	manual
4	Romanian	newspapers	written	5 million	manual
5	Russian	various	spoken & written	300 million	exhaustive
6	Swahili	news texts	written	12.5 million	exhaustive
7	Turkish	newspapers	written	20 million	manual

5. Predictions

The form-frequency correspondence principle in (5) above is stated in very general terms, and a number of specific testable predictions can be derived from it. We can test form-to-frequency predictions (overtly coded verbs should be less frequent) or frequency-to-form predictions (more frequent verbs should have less coding). And we can either look at individual verb pairs and individual languages, or aggregate the available cross-linguistic data, allowing us to quantify and compare general trends.

Linguists most often work on individual languages, and often they consider individual items. So we could test the individual-language prediction in (11):

(11) Prediction 1 (form-to-frequency, no aggregation)
In each language, in a causative verb pair, the causal member will be rarer than the noncausal member, while in an anticausative verb pair, the causal member will be more frequent than the noncausal member.

But at the level of individual verb pairs, there are not only many confirming cases (as in Table 3), but also many disconfirming cases (as in Table 4). The figures in these tables are absolute numbers from our corpus studies (see the Appendix for full data).

Table 3. Some verb pairs confirming Prediction 1

	causal	**noncausal**	**gloss**	**causal**	**noncausal**

		verb	verb		occurrences	occurrences
causatives (causal member expected to be rare)	Japanese	*kawakas-*	*kawak-*	'dry'	218	1578
	Russian	*kipjatit'*	*kipet'*	'boil'	514	5143
	Swahili	*gandisha*	*ganda*	'freeze'	20	82
	Turkish	*erit-*	*eri-*	'melt'	23	27
anticausatives (causal member expected to be frequent)	Maltese	*fetaħ*	*nfetaħ*	'open'	87	13
	Romanian	*închide*	*se închide*	'close'	40	10
	Russian	*raskolot'*	*raskolot'sja*	'split'	845	637
	Swahili	*vunja*	*vunjika*	'break'	883	376

Table 4. Some verb pairs disconfirming Prediction 1

		causal verb	noncausal verb	gloss	causal occurrences	noncausal occurrences
causatives (causal member expected to be rare)	Maltese	*tejjeb*	*tjieb*	'improve'	30	20
	Swahili	*poteza*	*potea*	'put/go out'	1728	654
	Turkish	*doldur-*	*dol-*	'fill'	38	12
anticausatives (causal member expected to be frequent)	Japanese	*tum(e)-*	*tumar-*	'fill'	953	1595
	Romanian	*usca*	*se usca*	'dry'	18	31
	Swahili	*pasua*	*pasuka*	'split'	105	252

Thus, if one is concerned primarily with the study of individual languages, it is easy not to see any regularities here, especially as there are some languages that have no confirming cases (such as English, which has no causative or anticausative pairs for our 20 verb meanings), and some languages that have more disconfirming than confirming cases (such as Romanian, as discussed below).

But recall that the form-frequency correspondence principle is stated as a tendency, and such tendencies are often easier to see by using aggregated data from a range of languages, which we can take to measure cross-linguistic trends. This has several advantages.

The advantage of aggregating coding types (causative vs. anticausative coding) across languages is that the effect of language-particular macro-types is eliminated. In individual languages, such macro-types can override the frequency effects. For example, Romanian codes 'dry' as an anticausative pair (*usca/se usca*, cf. Table 4), which goes against Prediction 1, but this clearly has to do with a language-particular macro-type: Romanian has a strong preference for coding causal-noncausal pairs by means of anticausatives using the morpheme *se*. Conversely, Turkish has a general preference for causative coding, and from this perspective it is not surprising that 'fill' (*doldur-/dol-,* cf. Table 4) is coded as a causative pair. Such language-specific macro-types are found in many domains of grammar, and they give each language its specific character or "genius". Table 5 shows that our seven languages differ considerably in their preference for causatives, anticausatives and others: The table ranks the languages by their *causative prominence*, i.e. the proportion of causatives among those verbs that occur in a causative or an anticausative pair (equipollents and labiles are disregarded for this measure).

Table 5: Different coding type trends in the seven languages

	CAUSATIVES	ANTICAUSATIVES	EQUIPOLLENTS	LABILES	% OF CAUSATIVES

Turkish	12	7	1	0	63
Japanese	8	7	4	1	53
Maltese	9	9	1	1	50
Swahili	6	8	7	0	43
Russian	1	13	6	0	7
Romanian	0	20	0	0	0
English	0	0	2	18	0

Macro-types interfere with the attempt to identify cross-linguistic frequency-based trends, but if we aggregate the coding types across languages, this gives us a measure of the general preference for causative and anticausative coding for each verb meaning. The aggregated data from 21 languages (from Haspelmath 1993: 104) are given in Table 6, which ranks our 20 verb meanings in (10) by causative prominence (again, this refers to the proportion of causatives among those verbs that occur in a causative or an anticausative pair, disregarding other kinds of pairs).

Table 6: Twenty verb meanings ranked by causative prominence[17]

RANK	MEANING	CAUSATIVES	ANTICAUSATIVES	% OF CAUSATIVES
1	boil	11.5	0.5	96
2	freeze	12	2	86
3	dry	10	3	77
4	wake up	9	3	75
5	go out/put out	7.5	3	71
6	sink	9.5	4	70
7	melt	10.5	5	68
8	stop	9	5.5	62
9	turn	7.5	8	48
10	burn	5	7	42
11	fill	5	8	38
12	rise/raise	4.5	12	27
13	improve	3	8.5	26
14	rock	4	12	25
15	connect	2.5	15	14
16	gather	2	15	12
17	open	1.5	13	10
18	break	1	12.5	7
19	close	1	15.5	6
20	split	0.5	11.5	4

(Data from 21 languages, from Haspelmath (1993). If a language has two counterpart verb pairs which behave differently, each counterpart pair counts 0.5, so that the numbers are not always integers.)

[17] Looking at the meanings of the verbs, we can make the intuitive generalization that the causative-prominent verbs express core-events that tend to occur spontaneously, while anticausative-prominent verbs express core-events that usually occur due to an external cause. In Haspelmath (1993: 105), a "spontaneity" scale was set up for verb meanings, which includes other meanings as well (see also Letuchiy 2010). However, here we do not invoke a semantic notion of spontaneity and explain the form asymmetries purely on the basis of frequency of occurrence (see Heidinger 2012, 2013+ for a different approach, where semantic spontaneity is taken as basic). The measure by which the verb meanings are ranked in Table 6 is called *causative prominence*, which is not a semantic notion. (Alternatively, one could say that causative prominence is a measure of spontaneity, as is done in Samardžić & Merlo (2012), but since "spontaneity" sounds like a semantically defined concept, this can give rise to misunderstandings.)

We thus see that different core-events are associated with quite different coding preferences. Some core-event meanings are very causative-prominent, others (such as 'split', 'close', 'break') are very anticausative-prominent, while still others occupy the middle ground.

In addition to aggregating the coding types, we can also aggregate the corpus frequencies. One advantage of this is that in this way we can measure the general likelihood that a causal or a noncausal verb will be used for a given core-event, regardless of its coding. After all, a skeptic could reject the frequency-causes-form explanation (of (9) above) and claim that the direction of causation is different: Longer forms are rarer because they are longer, and it is more economical to use shorter forms, so these tend to be more frequent. This reasoning is not absurd,[18] but if we aggregate the frequencies across languages, we reduce the possible effect from this interfering factor. Another advantage, which is probably even more important, is that the frequency-causes-form explanation really requires generalized frequencies, rather than the frequencies of the items whose form is to be explained. As we saw at the end of §3, the mechanism that creates short forms for frequent expressions is a lengthy process of diachronic change. The frequencies that are responsible for the current patterns of languages are not the current frequencies, but the frequencies at the time when the current patterns were created. But we cannot measure the earlier frequencies, so the only way to proceed is to use generalized frequencies (as was already done by Greenberg 1966, where frequency counts from Sanskrit and Latin were used to explain world-wide trends). The aggregated corpus frequencies will be presented in §7 below. On the analogy of the notion of causative prominence for meanings which are often expressed by causative pairs, we can speak of *(non)causal prominence* for verb meanings in which the (non)causal use is frequent.

With these two aggregated measures, we can make three further predictions:

(12) Prediction 2 (form-to-frequency, only form aggregated)
In verb meanings which are often expressed as causative pairs (verb meanings with high causative prominence, high on Table 6), the causal member will be rarer than the noncausal member in each language, while in verb meanings which are often expressed as anticausative pairs, the causal member will be more frequent.

(13) Prediction 3 (frequency-to-form, only frequency aggregated)
Verb meanings in which the causal member is rarer (verb meanings with low causal prominence, see Table 9 below) will tend to be coded as causative pairs in each language, while verb meanings in which the causal member is more frequent will tend to be expressed as anticausative pairs.

(14) Prediction 4 (both form and meaning aggregated)
Verb meanings with high causative prominence will exhibit low causal prominence, i.e. in verb meanings which are often coded as causatives across languages, the causal member will tend to be rare across languages. And conversely, verb meanings with low causative prominence will exhibit high causal prominence.

These predictions will be discussed further in §7 below, where we present the full results from our corpus studies.

6. Corpora and coding decisions

[18] In fact, there are probably some domains of language where coding length is indeed the cause of frequency (or rather rarity). A case in point is number: The reason why '100' is more frequent than '99' or '101' may well have to do with its relative shortness.

Before moving on to discuss the results obtained, it is worth outlining some of the decisions which needed to be made in order to obtain the frequencies in the case of the manual counts. The decisions fall into two categories: (i) decisions regarding the form of the verbs involved, and (ii) decisions regarding the meaning of the forms identified. We discuss each type in turn.

First, we limited ourselves to verbal forms[19] and excluded adjectival or nominal uses of our intended verbs (see examples in 15a-d). The reason for this was that nominalized and adjectival forms are ambiguous with respect to whether the verb is used causatively or not.

(15)

a. English *[..]; each morning she asked for **boiled** rather than scrambled eggs for breakfast*. (BNC)

b. Maltese *[Il-bozza] ilha diġà ġimagħtejn tkun mixgħula matul il-ġurnata u **mitfija** bil-lejl.* 'For the last two weeks [the bulb] was on during the day and **off (= put out)** at night.' (MLRS)

c. Romanian *Pentru as se evita **crăparea** peretelui, este preferată folosirea holtzsuruburilor*. 'In order to avoid [lit.] the **splitting** of the wall, it is preferred the use of the metal screws.' (RCNA)

d. Turkish *Ankara´da **toplanan** 150 bin işçi hükümeti ve siyasileri uyardı.* '150.000 workers, **gathered** in Ankara, warned the government and the politicians.' (Milliyet)

Passive uses of the verbs were generally included and coded as causative uses, as exemplified in (16a-c), since the situation is presented as having been caused by an external agent (and sometimes this agent is also coded).

(16)

a. English *The great library of Alexandria was **burnt** by Christians; in A.D. 411*. (BNC)

b. Maltese *Kull ammont dovut għandu **jinġabar** mill-awtorità kompetenti.* 'Any amounts due must be **gathered/collected** by the authorized authority.' (MLRS)

c. Romanian *De la Bucureşti, cele două tone de cartoane de joc care se adună săptămânal sânt aduse la distrus şi **topit***. 'From Bucharest, the two play cartons that are rounded up weekly are brought to be destroyed and **melted**.' (RCNA)

d. Turkish *Gemi İstanbul Boğazı´ndan geçerken **durduruldu***. 'Ship was **stopped** while passing the Bosphorus.*' (Milliyet)*

While in many cases, there is a formal distinction between the passive form of a verb and its adjectival form in a construction following a copula, e.g., compare *The door was opened* (passive, the door was opened by someone) vs. *The door was open* (attributive construction involving the copula 'be' and the adjective 'open'), this is not always the case for every verb in our sample. Cases where the two constructions were difficult to disentangle due to the lack of a morphological distinction, see the range of examples in (17a-c), were excluded from the

[19] In the languages that we studied, there was no need to consider periphrastic causatives, because all have nonperiphrastic causatives for all of the 20 verb meanings. Periphrastic causatives might occur as well (e.g. *make sth. break* in English), but we found that these are extremely rare, so we disregarded them. Likewise, we disregarded Japanese *-ase* causatives where these occur side by side with the ordinary causatives. But because of their rarity, including them would not have affected the results.

analysis, as were any other examples where it was difficult to ascertain whether the verb was used causally or noncausally.

(17)

a. English	*On this occasion I was **frozen** by anger and fear.* (BNC)
b. Romanian	*Şi fiind seară, în ziua aceea, întâia a săptămânii (duminică), şi uşile fiind **încuiate**, unde erau **adunaţi** ucenicii de frica iudeilor, a venit Iisus şi a stat în mijloc şi le-a zis: Pace vouă!* 'And being evening, on that day, the first of the week (Sunday), and the doors being **closed**, where the workers were **gathered** due to the fear of the Jews, Jesus came and sat in the middle and said to them: Peace onto you!' (RCNA)
c. Turkish	*[..]; dün **toplanan** komisyona ifade verdi.* '[He] declared to the commission **gathered** yesterday'. (*Milliyet*)

This brings us to the second type of decision made, namely one regarding meaning. We only considered examples where the verb under investigation was used with NP arguments (i.e., not including sentences like: *We **stopped** wanting a new car*). Our goal was to code first and foremost uses of the verbs which preserve the physical aspect of the action, that is, (physical) boiling, (physical) burning and so on, not idiomatic or metaphorical extensions of such uses.

However, as it is next to impossible to draw a neat line between concrete, physical uses of a verb and its idiomatic extensions, we included the latter, so long as they were extensions of the original meaning, rather than uses which no longer preserved any aspect of the physical sense. We also thought that the metaphorical uses of the verbs might tell us something about how the verb is used or conceptualised most saliently in the speakers' minds: causally or non-causally, depending on which type of use would be increasingly extended via metaphor. Thus, any new meanings that were not transparently related to this physical sense were excluded from the analysis, hence we allowed examples of the type in (18a-e), but not of those in (19a-e).

(18)

a. English	*But, if an error of judgement or a bad decision has been made, the vital thing is to recognise that, admit it and take immediate action to **break** the chain of events while a safe course of action is still possible.* (BNC)
b. English	*The words took time to **sink** in[20]; to herself as much as to the rest.* (BNC)
c. Maltese	*[...] **ħaraqni** meta bagħatli SMS bil-mistoqsija.* 'He burnt (= angered) me when he texted me the question.' (MLRS)
d. Romanian	*Nu bea, nu fuma, femeile **se topeau** după el.* '[He] didn't drink, didn't smoke, the women **melted** after him'.(RCNA)
e. Turkish	*Gençlerbirliği her nedense bu skordan sonra **uyandı**.* 'Gençlerbirliği **woke up** after this score for some reason'. *(Milliyet)*

[20] A further difficulty we encountered was how to treat complex verbs made up of [verb + particle], such as *sink in, boil up, melt away, dry up.* As discussed above, we included any uses which preserved (at least some of) the original meaning of the verb under scrutiny, and excluded those whose meanings diverged altogether from it (e.g., *rock on* meaning 'to party' or signalling approval).

(19)

a. English	*And then Woodleigh's secretary even came out last week, so I **gather**.* (BNC)
b. Maltese	*[…] jekk hemmx xi ħsieb li **jitwaqqaf** business centre fit-teknoloġija tal-informatika u l-innovazzjoni f'Għawdex.* […] whether they plan to **stop (= found, set up)** a business centre for information technology and innovation in Gozo. (MLRS)
c. Romanian	*Bogdan **s-a legat** cu lanţuri de piciorul statuii lui Mihail Kogălnicianu din Capitală.* 'Bogdan **tied** himself up (*se lega* also means '**connect**') with chains to the leg of the statue of Mihail Kogălnicianu in the capital [city].' (RCNA)
d. Romanian	*Ilinca **închide** telefonul*. 'Ilinca **puts down** (*închide* also means '**close**') the phone.'
e. Turkish	*O son rekorunu da giderken **kırdı.*** 'He **broke** his last record while leaving.' (Milliyet)

However, it seems quite likely that alternative decisions would not have affected the results much. Note that for the three languages that were not coded manually (Japanese, Russian, Swahili), we did include all occurrences of the verbs, regardless of their readings.

Our coding decisions reflect our desire to include as much of the data in the corpus as possible. We felt that such a conservative approach would better mirror the frequency tendencies of the verbs investigated. We limited ourselves to 50 uses of each verb once we convinced ourselves that appealing to 100 uses of 10 of our 20 verbs in Romanian and Maltese gave the same results (hence some of the results in the Appendix for Romanian and Maltese add up to 100 examples, whereas others to 50 examples).

7. Results

In this section, we present the results of our corpus study, the tests of the predictions of §5. The full data for all 140 verbs (twenty verbs from each of seven languages) with verb forms, coding type and frequencies extracted are presented in the Appendix.

Let us first look at Prediction 1, repeated here from §5:

(11) Prediction 1 (form-to-frequency, no aggregation)
In each language, in a causative verb pair, the causal member will be rarer than the noncausal member, while in an anticausative verb pair, the causal member will be more frequent than the noncausal member.

In the last column of the Tables A1-A7 in the Appendix, we have indicated for each verb pair whether it matches Prediction 1 or not. Table 7 summarizes the data.

Table 7. Number of verb pairs confirming and disconfirming Prediction 1

	confirming	disconfirming	not relevant
English	0	0	20
Japanese	11	5	4
Maltese	15	3	2
Romanian	9	11	0
Russian	10	4	6
Swahili	12	2	6
Turkish	12	7	1
Total	69	32	39

We see that a clear majority of verb pairs conform to the prediction. Independently of our work, Heidinger (2012) tested Prediction 1 for 16 verb pairs in a large French corpus (including a total of 3347 occurrences) and found it strongly confirmed. In addition, Narrog (2007) tested the prediction for 224 Old Japanese vrb pairs, and Narrog & Pardeshi (2013+) tested it for over 300 Modern Japanese verb pairs and found it confirmed as well.

This is encouraging, but recall from §5 that there is no direct causal link from language-particular frequencies to language-particular coding types. On the basis of our diachronic explanation in (9), we do not actually want to make the very strong Prediction 1. Due to the existence of macro-types, we find many disconfirming verbs, such as those listed in Table 4 above. And Romanian and English are two languages that do not show language-internal evidence for the hypothesis.

The macro-types are a strong interfering factor also for Prediction 3, where frequency is aggregated, but form is not. For this reason, we do not test this prediction in this paper. (The interested reader can easily work out to what extent it is fulfilled.) Instead, in the following we focus on Predictions 2 and 4. Prediction 2 is repeated below.

(12) Prediction 2 (form-to-frequency, only form aggregated)
In verb meanings which are often expressed as causative pairs (verb meanings with high causative prominence, high on Table 6), the causal member will be rarer than the noncausal member in each language, while in verb meanings which are often expressed as anticausative pairs, the causal member will be more frequent.

Figures 1(a-g)[21] are plots for each language that show the percentage of noncausal verbs for each of the 20 verb pairs of our verb sample. We see that in six of the seven languages, there is a tendency for verb pairs with higher causative prominence to occur more frequently as noncausals: The trendline is going up in all seven languages.

[21] All graphs and analyses were obtained using R (R Development Core Team 2004).

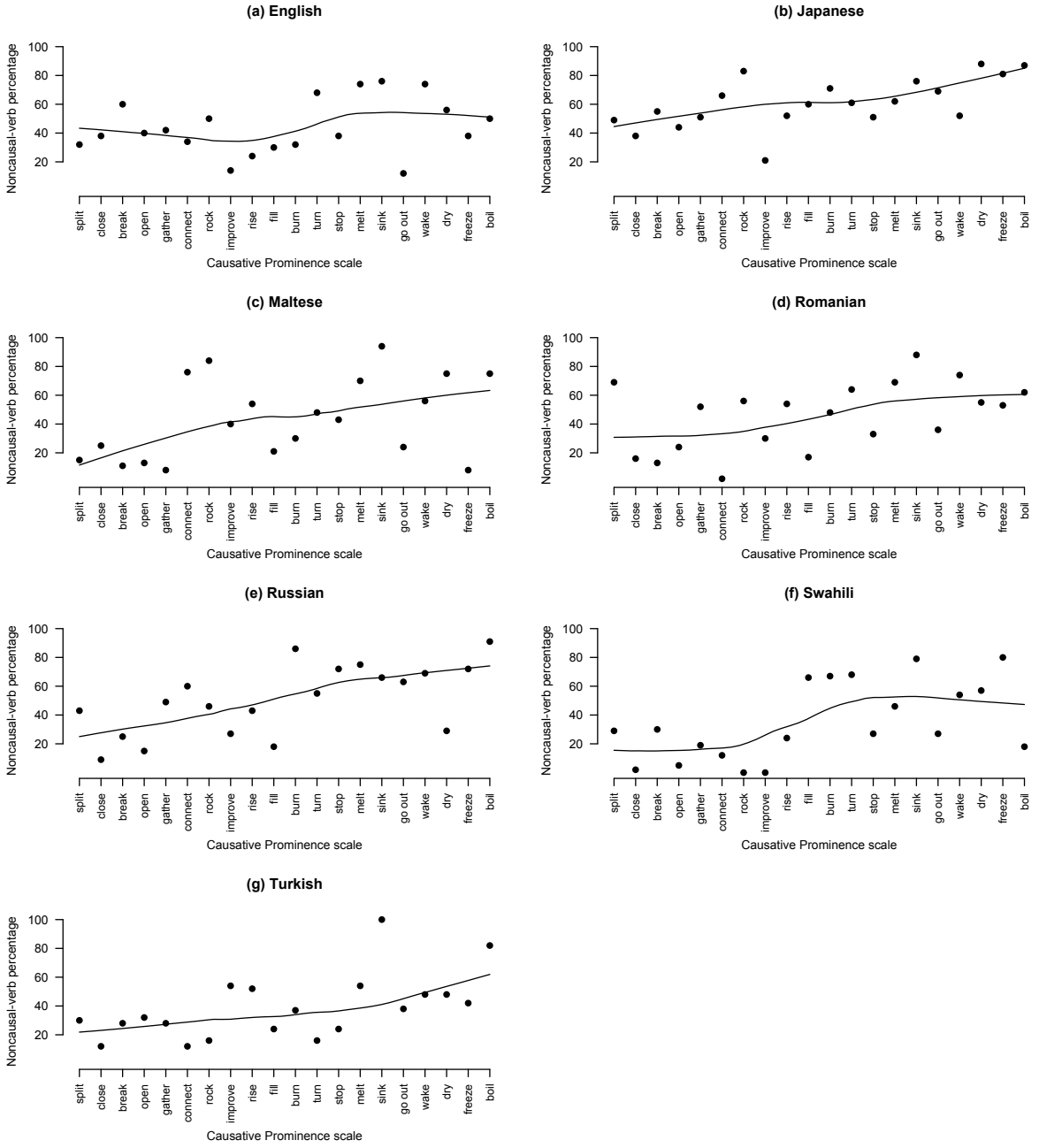

Figure 1(a-g). Smoothed scatter plots of the causalness ratio for each language (a-g). The x-axis shows the ordering of the twenty verbs, from least causative prominent to most causative prominent (cf. Table 6). The y-axis shows the noncausal prominence, i.e. the ratio of noncausal use versus total use of each verb (= noncausal / (causal + noncausal)).

The plots show that there is only one trendline where there are noticeable deviations from the general pattern observed, namely English.

In order to see whether the trends observed were statistically significant, we performed a number of tests, as detailed below. We used the Kendall Tau (τ) Rank test to check whether the trend lines observed for our languages are similarly ordered or not. This non-parametric test provides a Tau value (τ), which is between -1 and +1 (with -1 signalling a decreasing trend, and +1 signalling an increasing trend, and 0 signalling an absence of a trend), and an associated p-value outlining how sure we can be of its significance. The results are given in Table 8.

Table 8. Kendall Tau Rank tests for the seven languages and their mean

Language	Kendall Tau (τ)	p-value
English°	0.182	0.282
Japanese	0.508	0.002*
Maltese	0.286	0.085
Romanian	0.354	0.032*
Russian	0.487	0.003*
Turkish	0.396	0.018*
Swahili	0.370	0.025*
Mean	0.582	<0.001*

° In order to be able to calculate the Kendall Tau value, we filled in one missing value in our twenty items from means across the other languages in our sample for English *go out/put out*.

All of our seven languages give positive Tau values, which are generally associated with highly significant p-values (exceptions are Maltese, where the p-value is borderline significant, and English). However, if we exclude the verb 'freeze' in Maltese, τ=0.404, p=0.017. ('Freeze' is also exceptional in Samardžić & Merlo's (2012) findings.)

We also tested the mean[22] values for the twenty verbs investigated across our six languages and it has a Tau score of 0.747 indicating a strong positive trend, and a very significant associated p-value.

Next, we performed a Principal Component Analysis in order to identify patterns in the data and to see how many dimensions would be required in order to capture the data, without losing much information. The Principal Component Analysis shows that the first principal component explains 48% of the variation and it is the only one with an eigenvector greater than one. This means that one dimension is sufficient to capture the data, while the other factors essentially amount to random variation. Further support of this comes from the fact that the first principal component factor is 70% correlated to the mean of the language values for the twenty verbs ($p<0.0001$).

Thus, the corpus data from the seven languages support Prediction 2 to a very large extent. Independently of the present study, Samardžić & Merlo (2012) tested Prediction 2 for English on the basis of a much larger corpus (the Europarl corpus, which contains 1.5 million sentences) and found it very strongly confirmed for their data (correlation score $r = 0.77$, $p < 0.01$).

Next we turn to Prediction 4, again repeated here.

(11) Prediction 4 (both form and meaning aggregated)

> Verb meanings with high causative prominence will exhibit low causal prominence, i.e. in verb meanings which are often coded as causatives across languages, the causal member will tend to be rare across languages. And conversely, verb meanings with low causative prominence will exhibit high causal prominence.

To test this prediction, we had to aggregate the frequencies. We did this by averaging the percentages of causal uses across our seven languages.[23] Table 9 shows the twenty meanings ranked by percentage of noncausal uses (from highest to lowest).

[22] A Kendall Tau Rank test correlation of each language against each other shows that all languages are positively inter-correlated (with the exception of one single pair, between Swahili and Maltese, however, this has a very high associated p-value, so it is rather unstable). More significantly, a Kendall Tau Rank test correlation of each language against the mean of the remaining six languages shows that all six resulting correlations have positive Tau values (ranging from 0.254 to 0.550), and all but the lowest tau score (between Swahili and the remaining languages) have significant p-values.

Table 9: Twenty verb meanings ranked by average noncausal prominence (= average percentage of frequency of the noncausal member)

RANK	MEANING	AVE %	ENGLISH	JAPANESE	MALTESE	ROMANIAN	RUSSIAN	SWAHILI	TURKISH
1	sink	**83**	76	76	94	88	66	79	100
2	boil	**66**	50	87	75	62	91	18	82
3	melt	**64**	74	62	70	69	75	46	54
4	wake	**61**	74	52	56	74	69	64	48
5	dry	**58**	56	88	75	55	29	57	48
6	turn	**54**	68	61	48	64	55	68	16
7	freeze	**53**	38	81	8	53	72	80	42
8	burn	**53**	32	71	30	48	86	67	37
9	rock	**48**	50	83	84	56	46	0	16
10	rise	**43**	24	52	54	54	43	24	52
11	go out	**43**	(43)*	68	24	36	63	27	38
12	stop	**41**	38	51	43	33	72	27	24
13	split	**38**	32	49	15	69	43	29	30
14	connect	**37**	34	66	76	2	60	12	12
15	gather	**36**	42	51	8	52	49	19	28
16	fill	**34**	30	60	21	17	18	66	24
17	break	**32**	60	55	11	13	25	30	28
18	improve	**27**	14	21	40	30	27	0	54
19	open	**25**	40	44	13	24	15	5	32
20	close	**20**	38	38	25	16	9	2	12

* This figure is calculated by averaging the percentages in the other six languages, because we could not obtain a figure of use for *go out* in English.

We also checked at this point that the average percentages still agreed significantly with each of the language-internal rankings in order to make sure that no one language deviated significantly from the average (all Kendall Tau tests between the average frequency scale and the individual languages had at least $\tau > 0.390$, $p < 0.019$).

Next, the noncausal prominence scale obtained was compared to our causative prominence scale. Figure 2 gives the graphical representation of these two scales, showing that there is a close match between them, i.e., verbs which scored high on the average frequency-of-use scale also scored high on the causative prominence scale. As before, we verified the statistical significance of this correlation with the Kendall Tau test, which gives a highly significant result: $\tau = 0.653$, $p < 0.001$.

[23] Another way of aggregating causal prominence would be by averaging the rank scores for each verb across the seven languages. This yields very similar results.

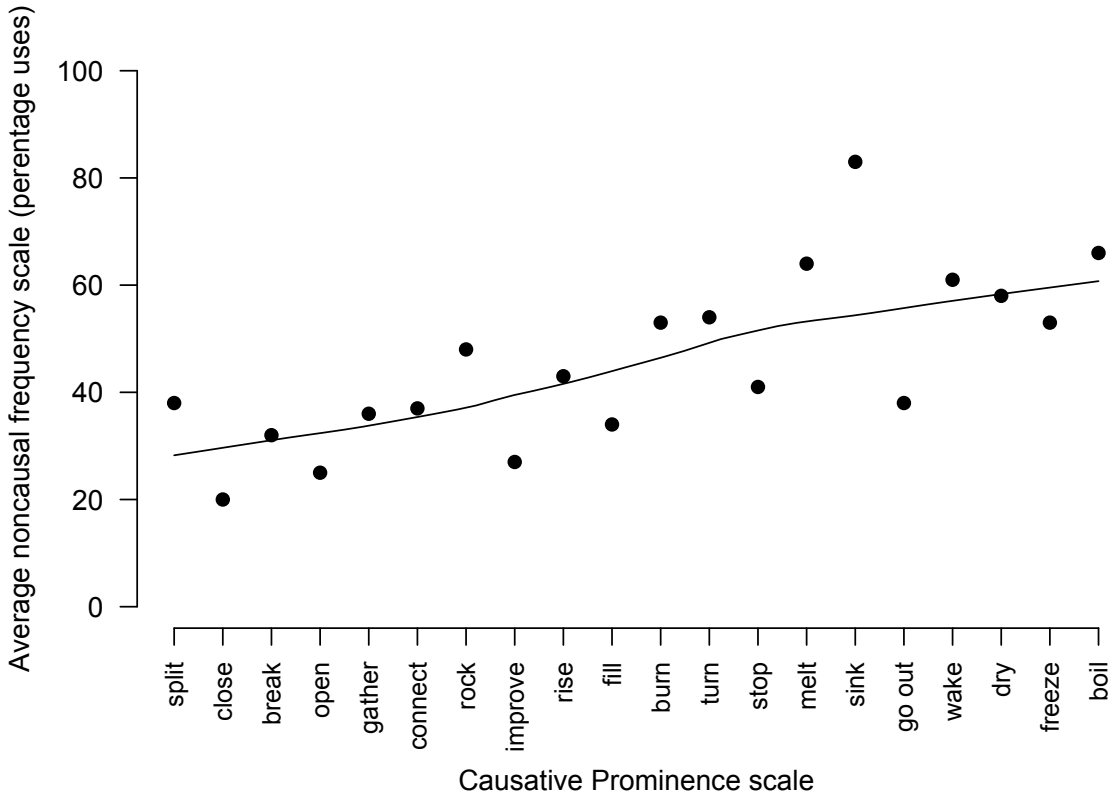

Figure 2. Smoothed scatter plot of the relationship between the average frequency scale and the causal prominence scale (average causal-verb proportion). The x-axis shows the ordering of the twenty verbs, from least causative prominent to most causative prominent. The y-axis shows the average frequency of use (percentage) of the noncausal member (average noncausalness) in the corpora.

Thus, Prediction 4 is very strongly confirmed as well.

We conclude that form-frequency correspondence is a strong effect in the seven languages and the 140 verb pairs that we examined, and we take it to support our explanation in terms of frequency, predictability and efficient coding.

8. Partial vs. full explanation

In addition to showing clear cross-linguistic trends, our frequency data of §7 also show a lot of "noise", i.e. unpredicted deviations from the expectations. Linguists are used to explanations that cover 100% of the facts, rather than explanations that only cover part of the data. They usually ask: What explains the rest of the data?

Our answer is that we do not know, but neither did we expect to be able to explain all formal patterns in all languages with a single general principle. The formal coding of meanings by languages depends on a wide variety of factors, including many historical accidents. By and large, languages maintain a balance of coding efficiency and tend to use short forms for frequent expressions and longer forms for rarer expressions. But this coding efficiency comes about as the cumulative effect of a highly diverse set of diachronic adaptive changes, not by any kind of system necessity. Deviations may arise for a variety of reasons, especially semantic change, which is independent of the form of the expression. A word with a specialized meaning may acquire a general meaning and thus become highly frequent without losing its longer form right away (or vice versa, acquire a specialized meaning and

become rarer). Examples are the English words *information*, *development*, and *particularly*, which are nowadays extremely frequent (at least in formal written language), but have a fairly long form, reflecting their earlier much more specialized meanings. Another factor is cultural change: Words like *yoke* or *louse* are nowadays rare, but have not become longer. Frequency distributions can change much more quickly than forms of words. Over time, the balance is likely to be re-established, but speakers are very conservative with respect to language form, so that form-frequency correspondence is not perfect.

Moreover, it should be kept in mind that the ranking of the 20 meanings of Table 1 in terms of (anti)causative prominence is tentative, as it is not based on a fully representative set of languages. We expect that if a more representative sample of languages is chosen to determine anticausative prominence, the amount of data that is predicted will increase (though it will never approach 100%).

A prominent work that differs from us in that it has the ambition to account for all cases in principle is Levin & Rappaport Hovav (1995). Let us briefly look at their (ultimately meaning-based) attempt to come to grips with the cross-linguistic diversity of causal-noncausal coding. They propose that *break* (intr.) is derived from *break* (tr.) by a rule that eliminates the causer in some way. Although they do not discuss languages other than English in any detail, they expect that the same is true for other languages, so they expect all change-of-state verbs to show anticausative coding if the alternation is overtly marked (and is not labile or equipollent).

In addition to 'break'-type verbs, they also discuss verbs like 'blossom' or 'decay', which cannot readily be used transitively in English, so they assume that these are fundamentally intransitive, and if they are to be used transitively, they need to undergo a causal-verb formation process. The contrast between 'break' and 'blossom' is ascribed to a difference in meaning: 'Break'-type verbs express *externally caused* events, while 'blossom'-type verbs express *internally caused* events, i.e. in the latter type of event, "some property inherent to the argument of the verb is "responsible" for bringing about the eventuality" (Levin & Rappaport Hovav 1995: 91). It is thus predicted in general that causative verb pairs should have an internally caused core-event, and anticausative verb pairs should have an externally caused core-event. If this is correct, then form-meaning parallelism can be maintained.

This is an interesting and very clear prediction, but does not match the cross-linguistic data, as causative pairs expressing externally caused events are not uncommon.[24] Levin & Rappaport Hovav are aware of this, but they still hope that their meaning-based account of English will carry over to other languages. They stress that the meaning of a verb needs to be examined in detail, and they suspect that counterpart verbs in different languages, as they are determined in coarse-grained typological studies like Haspelmath (1993), may actually differ with respect to the crucial parameter of internal vs. external causation. Thus, a verb that is roughly translated as 'melt' could express an externally caused melting event (as when something is caused to melt by heating it), or it could express an internally caused melting event (as when something melts without any external influence from heat). Only detailed language-particular analysis will tell whether it is internally or externally caused:

> "It is likely that this cross-linguistic variation arises because the meaning of a verb such as 'melt' is consistent with its describing either an internally or an externally caused eventuality. In fact, it should be possible to verify this prediction by looking at the range of subjects found with 'melt' in various languages; presumably, in languages where 'melt' is internally caused, it will only be found with ice or ice cream or other substances that melt at room temperature as its subject when intransitive (Levin & Rappaport Hovav 1995: 100)."

[24] Probably externally caused events are expressed more frequently as causatives than as anticausatives, because anticausatives are generally much less common than causatives, cf. Nichols et al. (2004: 162).

In Haspelmath's (1993) sample, there are ten languages with a causative pair for 'melt', so all these should be internally caused, and hence incompatible with substances that do not melt at room temperatures. Six of them are listed in (20).

(20)	French	*fondre/faire fondre*
	Finnish	*sulaa/sula-ttaa*
	Turkish	*eri-mek/eri-t-mek*
	Hindi-Urdu	*pighal-naa/pighl-aa-naa*
	Indonesian	*mencair/mencair-kan*
	Hungarian	*olvad/olvasz-t*
		'melt (intr.)/melt(tr.)'

At least in Finnish and Hungarian, it is possible to use the causative (*sula-ttaa, olvasz-t*) also with melting of wax, which requires a higher temperature (and thus some external causation), thus disconfirming the prediction.

While we agree with Levin & Rappaport Hovav that it is often very fruitful to seriously look for semantic determination of grammatical phenomena, we think that the enormous amount of cross-linguistic variation in the domain of causal-noncausal alternations makes it very unlikely that a semantic explanation will be found for all cases. In many languages, it is clear that there are different historical layers: for example, in Modern Greek, anticausative pairs often represent borrowings from Ancient Greek (e.g. *stréfo/stréfome* 'turn'), while native verbs tend to be labile (e.g. *jirízo* 'turn'), and in Modern German, old pairs going back to Proto-Germanic (such as *aufwachen/aufwecken* 'wake up') tend to be equipollent, while newly formed pairs tend to be anticausatives (Haspelmath 1993: 100). Thus, synchronic meaning is only one of the many factors determining the coding type of a particular verb.

9. Concluding remarks

We conclude that causative prominence of a causal-noncausal verb-meaning pair, as determined by studying form asymmetries across languages, correlates significantly with lower frequency of the causal member of the pair across languages. This is explained by our *form-frequency correspondence principle* in (5), which is itself explained by the tendency for languages to use efficient coding. When a long form becomes frequent, it tends to be shortened in language change, and when a short form becomes rare, it tends to become longer (Zipf 1935, Haspelmath 2008b).

As we noted in §3, this conclusion should not be surprising at all, because form-frequency correspondences are extremely widespread throughout the grammatical systems of languages. Linguists have traditionally had a tendency to explain form contrasts by meaning contrasts, but language structure is ultimately grounded in the communicative needs of speakers and hearers, and here predictability plays a very important role: Predictable meanings can be expressed in shorter ways, or can be omitted, and frequently expressed meanings are predictable. This kind of coding efficiency is reflected not only in the direction of morphological derivation, but also in periphrastic and syntactic patterns: The reason why we say *male nurse* is not that the concept of a male nurse is more complex than (or derived from) the concept of a (female) nurse, but that it's less frequent and less predictable. Similarly, we say *make someone laugh* (rather than *laugh someone*), but there is no need to say that the causation meaning is somehow different here from the causation meaning in *break something*: *make laugh* is simply much rarer (and much less expected) than *laugh*, so using an additional form is an efficient way to express these meanings.

One question that is often asked when someone gives a form-frequency correspondence explanation is what causes the frequency differences, and whether it could not be that a third factor causes both frequency asymmetries and form asymmetries. However, we are not aware of any concrete proposal of a third factor that would provide a serious explanation of the coding asymmetries.[25]

We offer no explanation for the frequency differences that we saw in Table 9 above, other than pointing out that it is intuitively plausible that events such as breaking and splitting which require a considerable input of force should be described more often in terms of a causer carrying out these actions, while natural events such as freezing, drying, and melting should be described more often as occurring spontaneously. These differences can be described in terms of "spontaneity" (cf. note 16), but since there is no independent way of measuring the spontaneity of an event, we do not regard this as an explanation. However, for our account, this is not necessary: Frequency asymmetries have diverse causes, but uniform consequences. Whatever the reason for the greater frequency of a form may be, it is bound to be shorter because of its higher predictability and the greater efficiency of a language system that exploits this predictability. This is illustrated by the causal chain in Figure 3, where we included a "factor X" that is responsible for frequency of use (but it plays no direct role in the explanation of language form).

factor X —> usage frequency —> predictability —> short form

Figure 3: Frequency causes predictability, which causes short form (see (9))

A possibility that we do not want to discount entirely is that predictability could be caused by something else (a factor X), which also causes frequency of use, as illustrated in Figure 4.

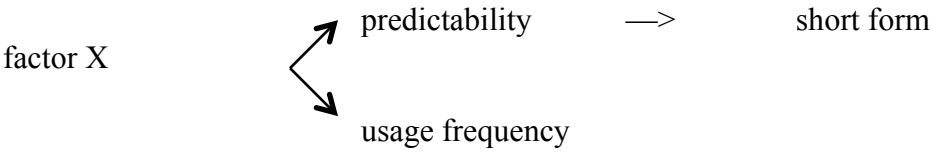

Figure 4: Some factor causes both predictability and frequency

Suppose factor X is real-world frequency: Since we know that drying happens more often spontaneously than under the influence of a causing agent, hearers can predict that speakers probably talk about noncausal drying and thus do not need a special anticausative marker to signal noncausal ‘dry’. This would be an example of a third factor causing both predictability (which causes short coding) and usage frequency. This would be compatible with the second part of Figure (3),[26] and crucially, even on this alternative scenario, there is no role for form-meaning parallelism. We do not know a good way of measuring predictability independently of usage frequency, so currently we propose the scenario in (9) and Figure 3, but we look forward to further research that might throw light on the issue.

[25] Heidinger (2012, 2013+) suggests that "spontaneity" is this third factor, but he describes no causal link between spontaneity and coding type.

[26] The particular case of ‘dry’ is not very plausible, however, because every drying process is of course caused by something. The difference between a towel drying “by itself“ and being dried by a machine is that in the first case, the cause is not salient at all, and we do not talk much about it (though one could say that the room temperature dries the towels). So if we want to avoid going back to usage frequency after all, we would have to invoke something like “cognitive-conceptual expectations“, as one reviewer puts it. The claim would be that a linguistic expression is more predictable because the hearer knows what the speaker thinks, independently of language.

Appendix: Data tables

In the data tables, we give the causal and noncausal members of each pair, then the number of causal and noncausal occurrences in th corpus (CAUS OCC, NONC OCC), then the coding type (Caus(ative), Anticaus(ative), Lab(ile), Equip(ollent); cf. Tables 1, 5), and finally the match with Prediction 1 of §5 ("y" for match, "–" for mismatch, and "(n.a.)" for irrelevant).

Table A1. English
(source: British National Corpus)

		CAUSAL VERB	NONCAUSAL VERB	CAUS OCC	NONC OCC	CODING TYPE	MATCH
1	boil	*boil*	*boil*	25	25	Lab	(n.a)
2	freeze	*freeze*	*freeze*	19	31	Lab	(n.a)
3	dry	*dry*	*dry*	28	22	Lab	(n.a)
4	wake up	*wake up*	*wake up*	37	13	Lab	(n.a)
5	put/go out	*put out*	*go out*	-	-	Equip	(n.a.)
6	sink	*sink*	*sink*	38	12	Lab	(n.a)
7	melt	*melt*	*melt*	37	13	Lab	(n.a)
8	stop	*stop*	*stop*	19	31	Lab	(n.a)
9	turn	*turn*	*turn*	34	16	Lab	(n.a)
10	burn	*burn*	*burn*	16	34	Lab	(n.a)
11	fill	*fill*	*fill*	15	35	Lab	(n.a)
12	raise/rise	*raise*	*rise*	12	38	Equip	(n.a)
13	improve	*improve*	*improve*	7	43	Lab	(n.a)
14	rock	*rock*	*rock*	25	25	Lab	(n.a)
15	connect	*connect*	*connect*	17	33	Lab	(n.a)
16	gather	*gather*	*gather*	21	29	Lab	(n.a)
17	open	*open*	*open*	20	30	Lab	(n.a)
18	break	*break*	*break*	30	20	Lab	(n.a)
19	close	*close*	*close*	19	31	Lab	(n.a)
20	split	*split*	*split*	16	34	Lab	(n.a)

Table A2. Japanese
(source: NINJAL-BCCWJ corpus, as of 2009, accessed through the NINJAL-LWP Project, by courtesy of Prashant Pardeshi)

		CAUSAL VERB	NONCAUSAL VERB	CAUS OCC	NONC OCC	CODING TYPE	MATCH
1	boil	*wakas-*	*wak-*	256	1847	Caus	y
2	freeze	*koor-ase-*	*koor-*	82	349	Caus	y
3	dry	*kawakas-*	*kawak-*	218	1578	Caus	y
4	wake	*okos-*	*oki-*	5162	5691	Caus	y
5	go /put out	*kes-*	*kie-*	2367	5056	Caus	y
6	sink	*sizume-*	*sizum-*	348	1172	Caus	y
7	melt	*tokas-*	*toke-*	450	725	Caus	y
8	stop	*tome-*	*tomar-*	5180	5477	Anticaus	–
9	turn/spin	*mawas-*	*mawar-*	2582	4052	Equip	(n.a)
10	burn	*moyas-*	*moe-*	549	1381	Caus	y
11	fill	*tume-*	*tumar-*	953	1595	Anticaus	–
12	rise/raise	*age-*	*agar-*	6092	6625	Anticaus	–
13	improve	*naos-*	*naor-*	1900	502	Equip	(n.a)
14	rock	*yuras-*	*yure-*	311	1509	Caus	y
15	connect	*tunag-*	*tunagar-*	1864	4313	Anticaus	–
16	gather	*atume-*	*atumar-*	3967	4117	Anticaus	–
17	open	*hirak-*	*hirak-*	4238	3363	Lab	(n.a)
18	break	*kowas-*	*koware-*	1044	1260	Equip	(n.a)
19	close	*sime- [sim-]*	*simar-*	972	606	Anticaus	y
20	split	*war-*	*ware-*	1310	1286	Anticaus	y

Table A3. Maltese
(source: Maltese Language Resource Server, MLRS Corpus)

		CAUSAL VERB	NONCAUSAL VERB	CAUS OCC	NONC OCC	CODING TYPE	MATCH
1	boil	*għalla*	*għala*	62	182	Caus	y
2	freeze	*ffriża*	*ffriża*	92	8	Lab	(n.a.)
3	dry	*nixxef*	*nixef*	24	73	Caus	y
4	wake up	*qajjem*	*qam*	22	78	Caus	y
5	put/go out	*tefa*	*ntefa*	38	12	Anticaus	y
6	sink	*għerreq*	*għereq*	3	47	Caus	y
7	melt	*dewweb*	*dab*	30	70	Caus	y
8	stop	*waqqaf*	*waqaf*	43	57	Caus	y
9	turn	*dawwar*	*dar*	26	24	Caus	–
10	burn	*ħaraq*	*nħaraq*	35	15	Anticaus	y
11	fill	*mela*	*mtela*	79	21	Anticaus	y
12	raise/rise	*għolla*	*għola*	23	27	Caus	y
13	improve	*tejjeb*	*tjieb*	30	20	Caus	–
14	rock	*bandal*	*tbandal*	8	42	Anticaus	–
15	connect	*għaqqad*	*ngħaqad*	12	38	Equip	(n.a.)
16	gather	*ġabar*	*nġabar*	46	4	Anticaus	y
17	open	*fetaħ*	*nfetaħ*	87	13	Anticaus	y
18	break	*kisser/kiser*	*nkiser/tkisser*	91	11	Anticaus	y
19	close	*għalaq*	*ngħalaq*	75	25	Anticaus	y
20	split	*qasam*	*nqasam*	85	15	Anticaus	y

Table A4. Romanian
(source: Romanian Corpus of Newspaper Articles, see Mihalcea & Năstase 2002)

		CAUSAL VERB	NONCAUSAL VERB	CAUS OCC	NONC OCC	CODING TYPE	MATCH
1	boil	*fierbe*	*(se) fierbe*	30	20	Anticaus	–
2	freeze	*îngheța*	*(se) îngheța*	26	24	Anticaus	–
3	dry	*usca*	*se usca*	31	18	Anticaus	–
4	wake up	*trezi*	*se trezi*	37	13	Anticaus	–
5	put/go out	*stinge*	*se stinge*	18	32	Anticaus	y
6	sink	*scufunda*	*se scufunda*	44	6	Anticaus	–
7	melt	*topi*	*se topi*	36	14	Anticaus	–
8	stop	*opri*	*(se) opri*	21	29	Anticaus	y
9	turn	*roti*	*se roti*	32	18	Anticaus	–
10	burn	*arde*	*(se) arde*	24	26	Anticaus	y
11	fill	*umple*	*se umple*	7	43	Anticaus	y
12	raise/rise	*ridica*	*se ridica*	27	23	Anticaus	–
13	improve	*îndrepta*	*se îndrepta*	15	35	Anticaus	y
14	rock	*legăna*	*se legăna*	28	22	Anticaus	–
15	connect	*lega*	*se lega*	2	48	Anticaus	y
16	gather	*aduna*	*se aduna*	26	24	Anticaus	–
17	open	*deschide*	*(se) deschide*	5	44	Anticaus	y
18	break	*sparge*	*se sparge*	10	40	Anticaus	y
19	close	*închide*	*(se) închide*	10	40	Anticaus	y
20	split	*crăpa*	*(se) crăpa*	30	20	Anticaus	–

Table A5. Russian
(source: National Corpus of Russian; only the nonderived aspct was considered)

		CAUSAL VERB	NONCAUSAL VERB	CAUS OCC	NONC OCC	CODING TYPE	MATCH
1	boil	*kipjatit'*	*kipet'*	514	5 143	Caus	y
2	freeze	*zamorozit'*	*zamerzut'*	1 229	3 171	Equip	(n.a.)
3	dry	*sušit'*	*soxnut'*	2 363	974	Equip	(n.a.)
4	wake up	*razbudit'*	*prosnut'sja*	5 843	12 835	Equip	(n.a.)
5	put/go out	*gasit'*	*gasnut'*	1 088	1 859	Equip	(n.a.)
6	sink	*utopit'*	*utonut'*	1 389	2 660	Equip	(n.a.)
7	melt	*plavit'*	*plavit'sja*	219	651	Anticaus	–
8	stop	*ostanovit'*	*ostanovit'sja*	13 998	36 694	Anticaus	–
9	turn	*povernut'*	*povernut'sja*	10 211	12 586	Anticaus	–
10	burn	*žeč'*	*goret'*	3 839	23 657	Equip	(n.a.)
11	fill	*napolnit'*	*napolnit'sja*	7 557	1 660	Anticaus	y
12	raise/rise	*podnjat'*	*podnjat'sja*	37 389	28 442	Anticaus	y
13	improve	*ulučšit'*	*ulučšit'sja*	2 400	877	Anticaus	y
14	rock	*kačat'*	*kačat'sja*	4 124	3 550	Anticaus	y
15	connect	*sočetat'*	*sočetat'sja*	1 433	2 153	Anticaus	–
16	gather	*sobrat'*	*sobrat'sja*	20 133	19 255	Anticaus	y
17	open	*otkryt'*	*otkryt'sja*	66 763	11 609	Anticaus	y
18	break	*lomat'*	*lomat'sja*	5 543	1 827	Anticaus	y
19	close	*zakryt'*	*zakryt'sja*	25 652	2 419	Anticaus	y
20	split	*raskolot'*	*raskolot'sja*	845	637	Anticaus	y

Table A6. Swahili
(source: Helsinki Corpus of Swahili)

		CAUSAL VERB	NONCAUSAL VERB	CAUS OCC	NONC OCC	CODING TYPE	MATCH
1	boil	*chemsha*	*chemka*	104	460	Equip	(n.a.)
2	freeze	*gandisha*	*ganda*	20	82	Caus	y
3	dry	*kausha*	*kauka*	152	201	Equip	(n.a.)
4	wake up	*amsha*	*amka*	324	381	Equip	(n.a.)
5	put/go out	*poteza*	*potea*	1728	654	Caus	–
6	sink	*zamisha*	*zama*	85	311	Caus	y
7	melt	*yeyusha*	*yeyuka*	102	88	Equip	y
8	stop	*maliza*	*malizika*	2376	900	Anticaus	y
9	turn	*geuza*	*geuka*	423	905	Equip	(n.a.)
10	burn	*unguza*	*ungua*	117	241	Caus	y
11	fill	*jaza*	*jaa*	456	892	Caus	y
12	raise/rise	*inua*	*inuka*	782	246	Anticaus	y
13	improve	*rekebisha*	*rekebika*	963	2	Equip	(n.a.)
14	rock	*zungusha*	*zunguka*	151	909	Equip	(n.a.)
15	connect	*unga*	*ungwa*	347	47	Anticaus	y
16	gather	*kusanya*	*kusanyika*	1225	283	Anticaus	y
17	open	*fungua*	*funguka*	2432	118	Anticaus	y
18	break	*vunja*	*vunjika*	883	376	Anticaus	y
19	close	*funga*	*fungika*	1369	22	Anticaus	y
20	split	*pasua*	*pasuka*	105	252	Anticaus	–

Table A7. Turkish
(source: Milliyet Newspaper Corpus, see Hakkani-Tür 2000)

		CAUSAL VERB	NONCAUSAL VERB	CAUS OCC	NONC OCC	CODING TYPE	MATCH
1	boil	*kaynat-*	*kayna-*	9	41	Caus	y
2	freeze	*don**dur**-*	*don-*	29	21	Caus	–
3	dry	*kuru**t**-*	*kuru-*	26	24	Caus	–
4	wake up	*uyan**dır**-*	*uyan-*	26	24	Caus	–
5	put/go out	*sön**dür**-*	*sön-*	31	19	Caus	–
6	sink	*bat**ır**-*	*bat-*	0	50	Caus	y
7	melt	*erit-*	*eri-*	23	27	Caus	y
8	stop	*dur**dur**-*	*dur-*	38	12	Caus	–
9	turn	*dön**dür**-*	*dön-*	42	8	Caus	–
10	burn	*ya**k**-*	*ya**n**-*	38	22	Equip	(n.a.)
11	fill	*doldur-*	*dol-*	38	12	Caus	–
12	raise/rise	*yükselt-*	*yüksel-*	24	26	Caus	y
13	improve	*geliş**tir**-*	*geliş-*	23	27	Caus	y
14	rock	*salla-*	*sallan-*	42	8	Anticaus	y
15	connect	*bağla-*	*bağlan-*	44	6	Anticaus	y
16	gather	*topla-*	*toplan-*	36	14	Anticaus	y
17	open	*aç-*	*açıl-*	34	16	Anticaus	y
18	break	*kır-*	*kırıl-*	36	14	Anticaus	y
19	close	*kapat-*	*kapan-*	44	6	Anticaus	y
20	split	*ayır-*	*ayrıl-*	35	15	Anticaus	y

Corpus sources - additional information

English	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/
Japanese	NINJAL-BCCWJ (http://www.ninjal.ac.jp/kotonoha/ex_8.html); as analyzed through the NINJAL-LWP (http://verbhandbook.ninjal.ac.jp/)
Maltese	MLRS, Maltese Language Resource Server, http://mlrs.research.um.edu.mt/
Romanian	http://www.cse.unt.edu/~rada/downloads.html#romanian, cf. Mihalcea & Năstase (2002)
Russian	National Corpus of Russian, URL: http://www.ruscorpora.ru/
Swahili	HCS 2004. Compilers: Institute for Asian and African Studies, University of Helsinki, and CSC – IT Center for Science.
Turkish	Milliyet Newspaper Corpus, provided by Kemal Oflazer, originally compiled by Dilek Hakkani-Tür and Gokhan Tür during the course of their PhD theses, cf. Hakkani-Tür (2000).

References

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, Elena Anagnostopoulou, and Florian Schäfer. 2006. The properties of anti-causatives crosslinguistically. In *Phases of interpretation*, ed. Mara Frascarelli, 187–211. Berlin: Mouton de Gruyter.

Borer, Hagit. 1991. The causative-inchoative alternation: a case study in parallel morphology. *The Linguistic Review* 8. 119–158.
Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733.
Clark, Eve V. & Clark, Herbert H. 1978. Universals, relativity, and language processing. In: Greenberg, Joseph H. (ed.) *Universals of human language. Vol. 1: Method and theory,* 225-277. Stanford: Stanford University Press.
Comrie, Bernard. 1985. Causative verb formation and other verb-derivational morphology. In Timothy Shopen (ed.), *Language Typology and Syntactic Description, Volume 3: Grammatical Categories and Lexicon*, 309–348. Cambridge: Cambridge University Press.
Croft, William. 1990. Possible verbs and the structure of events.. In S. Tsohatzidis (ed.), *Meaning and prototypes*, 48–73. London: Routledge.
Croft, William. 2003. *Typology and universals*. 2nd ed. Cambridge: Cambridge University Press.
Cysouw, Michael. 2008. Generalizing scales. In: Richards, Marc & Malchukov, Andrej L. (eds.) *Scales*, 379-396. (Linguistische Arbeitsberichte, 86.) Leipzig: Institut für Linguistik, University of Leipzig.
Dixon, Robert M. W. & Alexandra Y. Aikhenvald. 2000. Introduction. In Robert M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Changing valency: Case studies in transitivity*, 1–29. Cambridge: Cambridge University Press.
Doron, Edit. 2003. Agency and voice: The semantics of Semitic templates. *Natural Language Semantics* 11: 1–67.
Givón, T. 1991. Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language* 15(2). 335–370.
Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.
Guerssel, Mohamed & Hale, Kenneth & Laughren, Mary & Levin, Beth & White Eagle, Josie. 1985. A cross-linguistic study of transitivity alternations. *Chicago Linguistic Society* 21, part 2: 48-62.
Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.
Hakkani-Tür, Dilek Zeynep. 2000. *Statistical Language Modeling for Agglutinative Languages.* Ph.D. dissertation, Department of Computer Engineering, Bilkent University, Ankara.
Haspelmath, Martin. 1987. *Transitivity alternations of the anticausative type*. (Arbeitspapiere Des Instituts Für Sprachwissenschaft N.F. Nr. 4). Cologne: Universität zu Köln.
Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity*, 87-120. Amsterdam: Benjamins.
Haspelmath, Martin. 2004. Does linguistic explanation presuppose linguistic description? *Studies in Language* 28. 554–579.
Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(1). 25–70.
Haspelmath, Martin. 2008a. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.
Haspelmath, Martin. 2008b. Creating economical morphosyntactic patterns in language change. In Jeff Good (ed.), *Language universals and language change*, 185–214. Oxford: Oxford University Press.
Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3). 663–687.
Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(2). 31–80.
Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
Heidinger, Steffen. 2010. *French anticausatives: a diachronic perspective*. Berlin: De Gruyter.
Heidinger. Steffen. 2012. Frequenz und die Kodierung der Kausativ-Antikausativ-Alternation im Französischen. *Romanistisches Jahrbuch* 62. 31-58.
Heidinger, Steffen. 2013+. Spontaneity and the encoding of the causative–anticausative alternation in French and Spanish. Ms., University of Graz.
Koontz-Garboden, Andrew. 2009. Anticausativization. *Natural Language and Linguistic Theory* 27. 77-138.

Lehmann, Christian. 2007. On the upgrading of grammatical concepts. In Fons Moerdijk, Ariane van Santen & Rob Tempelaars (eds.), *Leven met woorden: Opstellen aangeboden aan Piet van Sterkenburg.*, 409–422. Leiden: Brill. (http://christianlehmann.eu/publ/upgrading.pdf).

Letuchiy, Alexander. 2010. Lability and spontaneity. In: Brandt, Patrick & Marco García García (eds.), *Transitivity: Form, Meaning, Acquisition, and Processing.* Amsterdam: Benjamins, 237-256.

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

Marantz, Alec. 1984. *On the nature of grammatical relations*. Cambridge, MA: MIT Press.

Mel'čuk, Igor' A. 1967. K ponjatiju slovoobrazovanija. *Izvestija Akademii Nauk SSSR, Serija literatury i jazyka* 26(4): 352-362.

Mihalcea, Rada & Vivi Năstase. 2002. Letter Level Learning for Language Independent Diacritics Restoration, in *Proceedings of the 6th Conference on Natural Language Learning* (CoNLL 2002), Taiwan.

Narrog, Heiko. 2007. Nihongo zita dōshi ni okeru yūhyōsei-sa no dōki-duke [On the motivation for markedness differences in Japanese transitivity verb pairs]. In: Kan Sasaki et al (eds) *Tadōsei no tsūgengoteki kenkyū – Tsunoda Tasaku Hakase kanreki kinen ronbunshū* [Cross-linguistic studies of transitivity – a festschrift for the 60th birthday of Doctor Tsunoda Tasaku], 295-306. Tokyo: Kuroshio Shuppan.

Narrog, Heiko & Prashant Pardeshi. 2013+. Frequency as the motivation for coding differences in Japanese transitivity pairs. Ms., Tohoku University & NINJAL Tokyo.

Nedjalkov, Vladimir P. & Georgij G. Sil'nickij. 1969. Tipologija morfologičeskogo i leksičeskogo kauzativov [Typology of morphological and lexical causatives]. In Alexander A. Xolodovič (ed.), *Tipologija kauzativnyx konstrukcij [Typology of causative constructions]*, 20–60. Leningrad: Nauka.

Nedyalkov, Vladimir P. & Georgy G. Silnitsky. 1973. The typology of morphological and lexical causatives. In Ferenc Kiefer (ed.), *Trends in Soviet theoretical linguistics*, 1–32. Dordrecht: Reidel.

Nichols, Johanna, David A. Peterson & Jonathan Barnes. 2004. Transitivizing and detransitivizing languages. *Linguistic Typology* 8(2). 149–211.

Payne, Thomas. 1997. *Describing morphosyntax: a guide for field linguists*. Cambridge: Cambridge University Press.

Piñón, Christopher. 2001. A finer look at the causative-inchoative alternation. In Rachel Hastings, Brendan Jackson, & Zsofia Zvolenszky (eds.), *Proceedings of semantics and linguistic theory XI*, vol. 11, 346–364. Ithaca, NY: CLC Publications.

Pylkkänen, Liina. 2008. *Introducing arguments*. Cambridge, Mass.: MIT Press.

R Development Core Team. 2004. R: A language and environment for statistical computing. On-line at http://www.R-project.org.

Rosch, Eleanor. 1978. Principles of categorization. In Eleanor Rosch & B. Lloyd (eds.), *Cognition and categorization*, 189–206. Hillsdale, NJ: Erlbaum.

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

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. *Language and Linguistics Compass* 3(2). 641–681.

Siewierska, Anna. 1984. *The passive : a comparative linguistic analysis*. London: Croom Helm.

Spagnol, Michael. 2011. *A tale of two morphologies: verb structure and argument alternations in Maltese.* Konstanz: University of Konstanz Ph.D. dissertation.

Wright, Saundra K. 2001. *Internally caused and externally caused change of state verbs*. Northwestern University Ph.D. dissertation.

Zipf, George K. 1935. *The psycho-biology of language*. Boston: Houghton Mifflin.