Semantic maps as metrics on meaning

Michael Cysouw
Max Planck Institute for Evolutionary Anthropology, Leipzig
cysouw@eva.mpg.de

Abstract

By using the world's linguistic diversity, the study of meaning can be transformed from an introspective inquiry into an subject of empirical investigation. For this to be possible, the notion of meaning has to be operationalised by defining the meaning of an expression as the collection of all contexts in which the expression can be used. Under this definition, meaning can be empirically investigated by sampling contexts. A semantic map is a technique to show the relations between such sampled contextual occurrences. Or, formulated more technically, a semantic map is a visualization of a metric on contexts sampled to represent a domain of meaning. Or, put more succinctly, a semantic map is a metric on meaning.

To establish such a metric, a notion of (dis)similarity is needed. The similarity between two meanings can be empirically investigated by looking at their encoding in many different languages. The more similar these encodings, in language after language, the more similar the contexts. So, to investigate the similarity between two contextualized meanings, only judgments about the similarity between expressions within the structure of individual languages are needed. As an example of the this approach, data on the cross-linguistic variation in inchoative/causative alternations from Haspelmath (1993) is reanalyzed.

Acknowledgments

I thank Martin Haspelmath and Caterina Mauri for helpful comments on how to improve the presentation of the somewhat tedious subject of this paper. Further, many of the concepts used in this paper arose in discussion with Bernhard Wälchli and should often just as well be considered his ideas (cf. Wälchli & Cysouw 2008). Every remaining inconsistency or unclarity is of course to blame completely on me.

1. Measuring meaning

Meaning is a particularly elusive property to measure. The central problem is that the meanings of linguistic expressions are variable across languages, and it is still mostly unknown how large this variability is. It does not really help to analyze the meaning of a language-specific expression (for example the English verb *to walk*) by saying that it expresses a general concept (like WALK). Such a change in typography still leaves open the question what the relation is to between WALK and, for example, the meaning of the German word *spazieren* or the Spanish word *andar*. Actually, without a more explicit definition of the concept WALK, asking whether *andar* expresses the concept WALK is not much different from asking whether *andar* means the same as *to walk*. Yet, individual linguistic expressions across languages never convey exactly the same range of senses, making such a simplistic approach to comparing meaning across languages devoid of content.

In this paper, I will defend the view that a much more profitable operationalization of the cross-linguistic variability of meaning is achieved by defining **the meaning of a language-specific expression as the collection of all contexts in which the expression can be used**. This definition represents, to some extent, a reversal of the intuitive notion of meaning. Meaning is typically thought of as some kind of property of a linguistic expression that governs its potential appearance in a particular context. In this conventional view, the main difficulty is how to express this property called "meaning". The approach to meaning proposed in this paper simply defines this property as the sum of all actual appearances. It is of course practically impossible to ever collect all appearances of a particular linguistic expression (be it a lexical or a grammatical item) in a living language—though it is possible for a dead language by including all documentation available—but samples of contexts can be used for any empirical question at hand (cf. Croft 2007; Wälchli & Cysouw 2008 for a similar approach to meaning).

Samples of the actual occurrences of expressions in concrete contexts can be used to compare the variation in meaning between different language-specific expressions. So, instead of assuming that we know what the English expression *walk* means, I propose to sample its meaning by considering various contextualized occurrences of walk-like situations. To compare expressions across languages, ideally the same sample of contexts should be used for all languages investigated. The parallel collection of such occurrences across languages can take various forms. It is possible to use extra-linguistic stimuli, like pictures (e.g. Levinson & Meira 2003) or video sequences (e.g. Majid et al. 2007), and investigate the linguistic expressions used to describe them. The contexts can also be defined purely linguistically, using descriptions of situations (e.g. Dahl 1985) or examples from parallel texts (e.g. Wälchli 2005).

In the practice of grammatical typology it is often impossible to collect sufficient parallel expressions because of the limited amount of material available, and because of the difficulty of finding native speakers for all the languages to be investigated. So, instead of concrete occurrences of language-specific expressions in context, normally somewhat larger domains of contexts are used in which an expression can occur (e.g. Haspelmath 1997). These domains are (more or less) explicitly defined 'chunks' of meaning, large enough to be identifiable from reference grammars, and small enough to capture the main distinctions of the cross-linguistic variation.[1] Both parallel expressions in context, as well as the somewhat more abstract domains of meaning as used conventionally in linguistic typology are called ANALYTICAL PRIMITIVES in Cysouw (2007).[2]

One of the consequences of comparing languages on the basis of an (empirical) selection of analytical primitives is that such a selection strongly reduces the range of possible meanings that can be identified across languages. Instead of the real-world continuous variation of possible meanings, a (finite) sample of analytical primitives only allows for a restricted, point-wise, granular view on this variation. In this approach, the meaning of a language-specific expression reduces to a subset of the sampled primitives. This subset consists of those sampled contexts in which the language-specific expression occurs. From the perspective of individual languages, the semantic analysis offered on the basis of such a selection of primitives might be somewhat coarse-grained, and perhaps to some extent even misleading. The most important gain of this approach, however, is that it offers a concrete operationalization of the cross-linguistic study of meaning. In this perspective, **the comparison of the meanings of two expressions from two different languages consists in the comparison of the selected subsets of analytical primitives.** Any deficits in the comparison arising from a biased selection of analytical primitives can easily be repaired by changing or extending the sample of primitives.

---

[1] It might be worthwhile to consider more precise definitions of such chunks of meaning as used in typology, for example using Natural Semantic Metalanguage (Wierzbicka 1996).

[2] The terms 'comparative concept' as used by Haspelmath (2008) and 'etic grid' as used by Levinson & Meira (2003: 487) are highly similar, if not identical, concepts to what I call 'analytical primitive'.

To be able to make cross-linguistic comparisons between language-specific expressions from different languages, first the internal structure among the primitives has to be considered. This paper deals with the empirical establishment of such structure among analytical primitives in the form of semantic maps (Section 2). The actual comparison of language-specific expressions (i.e. questions like 'how similar is English *walk* to Spanish *andar*, and in which aspects to they differ?) will not be further pursued here.[3] In a very general sense, the structure among analytical primitives amounts to establishing a **metric on analytic primitives, i.e. a specification of the distances (or "dissimilarities") between them**, as will be discussed in Section 3. One way to empirically arrive at these dissimilarities between primitives is to use the cross-linguistic diversity in the encoding of the primitives, as discussed in Section 4. **Only language-specific analysis is necessary to establish dissimilarities between primitives—there are not cross-linguistic judgements necessary.** This important insight led to the establishment of semantic maps in the first place, but will be generalized here in Section 5. In Section 6, I will argue that **both form and behavior can be analyzed as language-specific encoding**. An example of this conceptualization of the cross-linguistic study of meaning is presented in Section 7, reanalyzing data on the inchoative/causative alternation from Haspelmath (1993).

2. Semantic maps

Analytical primitives are not just points in an unstructured cloud of semantic space. Some primitives are more similar to each other than others. Such structure among analytical primitives is suitably analyzed by using semantic maps (cf. Haspelmath 2003). Semantic maps are a special kind of analysis and display of the internal structure of a sample of analytical primitives. My use of the terms SEMANTIC SPACE and SEMANTIC MAP is most closely related to Haspelmath's terminology, in which 'a semantic map is a geometrical representation of functions in "conceptual/semantic space" ' (Haspelmath 2003: 213). This is different from the terminology used by Croft (although there is no difference in content), who uses the term 'conceptual space' for the geometrical representation, and 'semantic map' for the language-specific instantiation (cf. Croft 2001: 92ff; Croft 2003: 133-139, #67202; Croft & Poole 2008: 3). The different terminologies are summarized in Table 1.

Differently from the received view of such semantic maps, I propose here to strictly separate the notion of a semantic map into two different aspects, namely the STRUCTURE among

[3] For some first attempts to compare the meaning of language-specific expressions, see Cysouw (2007) and Wälchli & Cysouw (2008).

the primitives and the DISPLAY of this structure. The structure itself will be formulated as a metric on the primitives; the display of the structure is the semantic map proper. Given a particular set of data, there will both be different ways to establish the structure among the primitives, and there will be different ways to display any structure attested. Because of the multitude of possibilities, it is particularly important to separate effects stemming from the decision on how to measure the structure from effects resulting from the specific method to visualize the structure. In this paper, I will only discuss approaches to the establishment of the structure among primitives. The discussion of the various possible visualizations is left to another occasion.

Table 1. Terminological clarification.

<table>
<tr><th rowspan="2">Concept</th><th colspan="3">Terminology</th></tr>
<tr><th>This paper</th><th>Haspelmath</th><th>Croft</th></tr>
<tr><td>Collection of all possible analytical primitives</td><td>conceptual/ semantic space</td><td>conceptual/ semantic space</td><td>–</td></tr>
<tr><td>Structure within the set of analytical primitives</td><td>cross-linguistic metric on meaning</td><td rowspan="2">semantic map</td><td rowspan="2">conceptual space</td></tr>
<tr><td>Graphical representation of attested structure</td><td>semantic map</td></tr>
<tr><td>Language-specific encoding of analytical primitives</td><td>language-specific metric on meaning</td><td rowspan="2">boundaries in semantic map</td><td rowspan="2">semantic map</td></tr>
<tr><td>Graphical representation of language-specific encoding</td><td>language map</td></tr>
</table>

3. Metrics and distance matrices

A METRIC is the mathematical explication of a notion of distance (or dissimilarity, i.e. the opposite of similarity). In our daily world, the most natural notion of distance is the Euclidean distance, i.e. the distance “as the crow flies”. However, when moving from point A to B it is often not possible to take the direct route (if you are not a crow), so another natural metric is the ground travel distance. This notion of distance can widely deviate from the straight-line Euclidean distance, namely when there is no (approximately) direct route to get from A to B while staying on the ground. Still another way to measure distance in daily life is to take the time it takes to get from A to B. Again, this notion of distance might give a rather different perspective on our surroundings depending on transportation possibilities. These different ways of measuring distance illustrate that **any notion of distance is a question of perspective, and is not in any sense pre-established by the nature of the objects investigated.** This holds also for metrics on meaning: what counts as similar in meaning depends on what perspective one wants to take.[4]

The result of applying a metric on some data is a table of pairwise distances for all pairs of objects investigated: a DISTANCE MATRIX. So, given some data and a decision on how to interpret the data (the metric), distances between pairs of objects can be computed. Normally, such pairwise distances are expressed as a (fractional) number between zero and one. At the one extreme "0" indicates "no distance", i.e. the two objects are the same, and at the other extreme "1" indicates "maximal distance", i.e. the objects are completely different. It is not necessary to normalize distances to this zero-one interval, but it makes it easier to combine distance matrices. Also, decimally written values between zero and one can intuitively be taken to represent percentages. For example, a distance of 0.54733 can be interpreted as "almost 55% of the maximal distance". And, finally, the distances between zero and one are easily switched to similarities, because when two objects have a distance of d, then they have a similarity of $1-d$.

Distance matrices can become bewilderingly large and difficult to interpret for a human being. For example, with only 10 analytical primitives there are already $10 \times 9 \div 2 = 45$ distances between pairs of primitives. Just looking at such a long list of numbers will mostly not result in very revealing insights, because it is difficult to discern meaningful distinctions among the wealth of available information. There are many ways to help a human being make sense of what would otherwise be categorized as information overload, but this is an extensive topic that I will not discuss in detail here. Suffice it to say that visualization is a highly powerful technique, though it can also be deceptive because human eyes (and brains) tend to see patterns also when there are none. For this reason it is advisable never to rely on just one visualization, and to always determine afterwards whether any patterns

[4] It is an open question whether different approaches to measuring meaning converge. If there exists something like "the" meaning, than this should be the case. Given the framework for investigating meaning as sketched in this paper, this question becomes an empirical problem.

perceived are really statistically significant. Finally, it is important to recognize that every visualization is always an abstraction of the underlying data, or, put more bluntly, many details are necessarily ignored, or intentionally misrepresented, in the process of making a visually pleasing graphic display. The network-like graph used for traditional semantic maps (cf. Haspelmath 2003) is such a pleasing graphic display, for which various fundamental abstractions of the available data are made (cf. Cysouw 2007 for a detailed criticism).

4. Using linguistic diversity

The basic intuition behind the semantic map approach to meaning is that **cross-linguistic variation in the expression of meaning can be used as a proxy to the investigation of meaning itself**. Concretely, recurrent similarity in form reflects similarity in meaning, or, as Haiman (1985: 19) puts it: “recurrent identity of form between different grammatical categories will always reflect some perceived similarity in communicative function.” Thus, the assumption is that, when the expression of two meanings are similar in language after language, then the two meanings themselves are similar. Individual languages might (and will) deviate from any general pattern, but when combining many languages, overall the cross-linguistic regularities will overshadow such aberrant cases.[5]

Formulated in the framework set up in the previous sections, this basic intuition can be formalized as follows. To start off, a sample of analytical primitives has to be established, and expressions of these primitives have to be collected for a sample of the world’s languages. Then, for each language individually, the similarity between these expressions can be established within the structure of the language (i.e. only language-specific constructions and language-internal form-similarities are investigated). Technically formulated, this means that a language-specific metric on the expressions will be set up—a different one for each language (see Section 7.2 for a concrete example of how this might work). Then, **the**

---

[5] This approach assumes that every meaning is expressible in all human languages. The expression of a meaning might be easier in some languages, and take more effort in others, but it is possible everywhere. However, there are various obvious complications with this assumption; see for example Levinson (2003) for a challenge to this assumption regarding the expression of spatial concepts. Further, I will ignore the complications arising from the fact that most languages will have many different ways to express a particular meaning. This is not problematic for the goal of computing meaning similarities, but the mathematical details will become a bit more involved.

cross-linguistic metric on the analytical primitives ("semantic map") is the average of the language-specific metrics on the expressions collected. This simple statement represents a big step forward for any empirical investigation of meaning (cf. Haspelmath 2003: 230-233). Instead of requiring elusive judgments about the similarities between meanings, all that is needed now are very concrete judgments about the similarity between language-specific expressions within one and the same language. So, to establish a cross-linguistically viable metric on meaning, it is not necessary to perform cross-linguistic comparisons of expressions from different languages. Purely on the basis of many language-specific analyses, it is possible to arrive at general results.

5. Constructions and strategies

To establish a metric on expressions, a notion of (dis)similarity between expressions is needed. There are basically two different kinds of (dis)similarity. The first possibility is to compare the amount of shared morphophonological material between expressions. Such similarity is purely language-specific and cannot be used to directly compare expressions across languages (except of course in historical-comparative reconstruction). In contrast, more abstract characteristics are necessary to establish the cross-linguistic similarity between expressions. Examples of such more abstract characteristics are the order of elements, the length of expressions, or the degree of fusion between elements (e.g. isolation, concatenation, or non-linear morphology). This is an important differentiation, as made implicitly in the semantic map literature. The first similarity leads to a LANGUAGE-SPECIFIC EXPRESSION METRIC ("constructions") and the second to a CROSS-LINGUISTIC EXPRESSION METRIC ("strategies"). Most of the comparisons in the field of linguistic typology is based on comparing cross-linguistic strategies (cf. Croft 2003: 31ff.). However, semantic maps are purely based on language-specific constructions.

Given a language-specific metric, a LANGUAGE-SPECIFIC CONSTRUCTION (in the sense of Croft 2001; Goldberg 2006) is a set of language-specific expressions that are highly similar from the perspective of the metric. What exactly means "highly similar" is of course less obvious, but any disputable similarity-boundary will likely be reflected by an equally vague notion of what defines the construction involved. Though different operationalization of similarity can be used (and see Section 7 for a few possibilities), I am strongly in favor of a gradient notion of language-specific constructions (i.e. individual expressions in a language are more or less similar on a continuous scale). I think it is misguided to look for any strict definition of constructions that discretely classifies all expressions of a language into separate constructions.

Being the counterpart to constructions, a TYPOLOGICAL STRATEGY is a set of expressions that are highly similar from the perspective of a cross-linguistic metric (the term “strategy”, now commonly found in the typological literature, was probably first used in this sense by Keenan & Comrie 1977: 64). Just like constructions are abstractions of language-specific metrics, strategies are abstractions of cross-linguistic metrics. For example, consider the causative/inchoative alternation, to be discussed extensively in Section 7. The English inchoative expression *the vessel is destroyed* has a causative counterpart *the torpedo destroyed the vessel*. Now, the language-specific construction to derive the anticausative from the causative in English for the verb *destroy* is to use an expression with the verb *to be*. From a cross-linguistic perspective, this alternation is an example of an ‘anticausative’ typological strategy, using the terminology of Haspelmath (1993: 91), because the inchoative is transparently derived from the causative.

The main claim of the semantic map approach is that **a metric on meaning (“semantic map”) can be established purely on the basis of many language-specific expression metrics (“constructions”), averaged over a diverse sample of languages.** Cross-linguistic metrics (“strategies”) are not necessary for this goal.[6]

6. Coding and behavior

There are many different possibilities to establish a language-specific expression metric. In the next section, concrete example of three different metrics on the same data are discussed in detail. One somewhat atypical aspect of the upcoming example is that the metrics are based on pairs of expressions, not on single expressions like in traditional semantic maps (Haspelmath 2003). The approach to consider the relation between two expressions is reminiscent of Keenan’s (1976: 306-307) ‘transformational behavior’. Following Keenan, the terms ‘coding’ and ‘behavior’ have become widespread for the analysis of grammatical relations. Generalizing this distinction, I will use the term ‘coding properties’ for properties of

[6] One auspicious prospect is that an association between a cross-linguistic metric (“strategy”) and a language-specific metric (“construction”) represents a generalization of what is known in linguistics as a “hierarchy” or a “scale”. Establishing such a correlation is not trivial because language-specific metrics cannot be compared directly across languages (see the example at the end of Section 7.2 for a first glimpse of this prospect, and see Cysouw 2008 for a more elaborate discussion).

individual expressions, while ‘behavioral properties’ are properties of the relation between expressions.

> “The properties may be pragmatic, semantic, or syntactic. And of the syntactic ones, some concern properties internal to a single sentence [i.e. ‘coding’, MC] and other concern the relation between a b-sentence and some modification of it [i.e. ‘behavior’, MC].” (Keenan 1976: 312)

Under this definition, the opposition coding vs. behavior is independent from the opposition construction vs. strategy, as discussed in the previous section. There are thus four logically possible combinations that represent different approaches to characterize and compare expressions.

First, a **coding strategy** is a cross-linguistic classification of the structure of a particular expression. This is the most prototypical kind of approach in linguistic typology. The classic example is the typology of relative clause structures, distinguishing types like ‘relative pronoun strategy’ or the ‘internally headed relative clauses’ (Lehmann 1984; Comrie & Kuteva 2005). Second, a **behavioral strategy** is a cross-linguistic classification of the relation between various expressions (typically two, but possibly more). A classic example is the relation between a regular matrix sentence like *John swept the floor* and the corresponding action nominal construction *John’s sweeping of the floor* (cf. Keenan 1976: 321). For this behavior, a cross-linguistic classification of possible strategies used by human languages has been developed by Koptjevskaya-Tamm (1993, 2005).

Third, **constructional coding** is a characterization of the language-specific form of a expression. This is the typical information that is used in traditional semantic maps. The more similar two expressions as to their constructional coding, the closer their meaning (when averaged over a large number of languages). Finally, **constructional behavior** is the fourth possibility. This possibility to characterize expressions to is not very widely acknowledged in the typological literature, but this will be the approach that I will use in the case study in the next section. The basic idea is to compare the combined language-specific forms of all alternative expressions that are relevant for the behavior.

7. Case study

7.1. Causative/inchoative alternations

As an example of the approach presented here I will reanalyze the data from Haspelmath (1993) on the causative/inchoative alternation. In his paper, Haspelmath addresses the question how languages mark the predicate in the alternation between an inchoative expression, like *the water boiled*, an a causative expression, like *the man boiled the water*. In the case of the English predicate *boil* there is no difference in the marking, but for other alternations, like *die/kill* or *be destroyed/destroy*, the difference between the inchoative and the causative version is reflected in the lexical or morphological form of the predicate. The approach of Haspelmath's study is to investigate cross-linguistic **strategies** of expressing the relation between inchoative and causative meanings, but that aspect of his study will not be the main focus of this paper (some preliminary hints on the relation between strategies and meaning will be given at the end of Section 7.2). Instead, I will investigate the **relations between the meanings of the predicates** by investigating the language-specific marking that is used to express the inchoative/causative alternation.

Haspelmath investigated the inchoative/causative alternation for 31 analytical primitives ("lexical meanings") in 21 languages. The 31 meanings investigated are repeated here in Table 2 (adapted from Table 2 in Haspelmath 1993: 97).[7] The translations of these meanings in all 21 languages are added as an appendix to Haspelmath's paper, allowing for the current reanalysis of the data.[8]

---

[7] The primitives used in this paper are a somewhat special kind of lexical meanings, because they are neutral with respect to the causative/inchoative alternation. For example, the English pair *kill/die* is considered to be a single primitive here, notwithstanding the lexical suppletion. It is important to realize that not all languages have suppletion for the same primitives, so cross-linguistically the pair *kill/die* has to treated equivalent to a non-suppletive pair like *destroy/be destroyed*.

[8] To simplify the calculations, I have maximally included one expression for each meaning in each language. In some cases, Haspelmath lists more than one possible expression, and in those cases I have semi-randomly chosen one of the options. If possible, I have discarded idiosyncratic alternations showing inchoative/causative morphology that was not found in any other sampled expressions of the same language. Only if all alternatives used constructions also found elsewhere did I randomly select one of them. This was only necessary in a handful of cases.

Table 2. Inchoative/causative pairs investigated in Haspelmath (1993).

No.	Inchoative	Causative	No.	Inchoative	Causative
1	wake up	wake up	17	connect	connect
2	break	break	18	boil	boil
3	burn	burn	19	rock	rock
4	die	kill	20	go out	put out
5	open	open	21	rise	raise
6	close	close	22	finish	finish
7	begin	begin	23	turn	turn
8	learn	teach	24	roll	roll
9	gather	gather	25	freeze	freeze
10	spread	spread	26	dissolve	dissolve
11	sink	sink	27	fill	fill
12	change	change	28	improve	improve
13	melt	melt	29	dry	dry
14	be destroyed	destroy	30	split	split
15	get lost	lose	31	stop	stop
16	develop	develop			

I will use the language-specific marking of the inchoative/causative alternation of the meanings listed in Table 2 as a proxy to the measurement of the similarity between the meanings. For example, the English expression of meaning 1, *wake up/wake up*, does not use any marking to differentiate inchoative from causative. This means that meaning 1 is somewhat alike to meaning 2, in English expressed as *break/break*, which likewise does not differentiate inchoative from causative. A similar situation is found in French. The French expressions of meanings 1 and 2 also use the same construction (viz. a reflexive pronoun with the inchoative: *se réveiller/réveiller* and *se briser/briser*, respectively). This is again an indication that these two meanings are somewhat alike. In German, though, meanings 1 and 2 do not use the same process (viz. an ablaut-like alternation in *aufwachen/aufwecken* vs. no differentiation in *zerbrechen/zerbrechen,* respectively), which is an indication that the meanings 1 and 2 are also somewhat different.

The marking of the inchoative/causative alternation on the predicate is just one of very many possible approaches to investigating similarity between meanings, or, to paraphrase a claim made in Section 3, any notion of similarity is a question of perspective, and is not in any sense pre-established by the nature of the expressions investigated. The rather abstract nature of the notion of similarity as used here (i.e the formation of the inchoative/causative alternation) is appealing because it allows for the comparison of otherwise difficult-to-compare meanings, like 'wake up' and 'break'.[9] In the following, I will discuss three different ways to operationalize this language-specific notion of similarity between expressions.

7.2. Metric A: Language-specific constructions

The first example of a language-specific similarity between expressions will be based on establishing language-specific constructions. I will here define a construction as a regular morphosyntactic relation between an inchoative and an causative verb form. Such relations are purely language-specific (see the Appendix for a complete survey of all constructions distinguished for this paper). For example, in English, the 31 meanings shown in Table 2 can be classified as belonging to seven language-specific constructions. There is one large class consisting of verbs that do not show any difference in morphology between inchoative and causative usage (viz. *wake up, break, burn, open,* etc.). The remaining six classes each consist only of one meaning, using different inchoative/causative alternations in each case (viz. *die/kill, learn/teach, be destroyed/destroy, get lost/lose, go out/put out,* and *rise/raise*). As an example, just the first three meanings are shown in Table 3, all three being marked as belonging to the same class (called "E-1", where the "E" indicates that this is a language-specific class for English only).

For other languages, these classifications will look different. For example, in French there are five different classes. First, there is one large class in which the inchoative form is marked with an reflexive pronoun (e.g. 1: *se réveiller/réveiller* and 2: *se briser/briser*). Second, there is another large class in which there is no difference between inchoative and causative

[9] Most theories of meaning will not have much to say about the relation between 'wake up' and 'break', other than coincidental points such as the observation that in English the metaphor *break of day* is used for the morning, which is also the prototypical time to wake up.

verb forms (e.g. 3: *brûler/brûler*). Then, there is a small class where the causative is formed by adding the verb *faire* (among the current 31 meaning this is found only for 13: *fondre/faire fondre* and 18: *bouillir/faire bouillir*). Finally, there are two French expressions that do not have any parallel among the current 31 meanings, so they make up their own class (viz. 4: *mourir/tuer* and 14: *être détruit/détruir*).

Table 3. Excerpt of language-specific classes for inchoative/causative alternations.

	English		French		German	
No.	Form	Class	Form	Class	Form	Class
1	*wake up/wake up*	E-1	*se réveiller/réveiller*	F-1	*aufwachen/aufwecken*	G-1
2	*break/break*	E-1	*se briser/briser*	F-1	*zerbrechen/zerbrechen*	G-2
3	*burn/burn*	E-1	*brûler/brûler*	F-2	*verbrennen/verbrennen*	G-2

Once established for all languages in the sample, these language-specific classes ("constructions") can now be used to calculate the (dis)similarity between the primitives ("lexical meanings"). Basically, every pair of meanings is considered separately for all 21 languages, and the number of languages is counted for which the two meanings belong to different constructions. The higher this number, the more languages put the meanings in different constructions, indicating that the meanings are different. For example, considering meanings 1 and 2 in the excerpt of the data shown in Table 3, these two meanings belong to the same class in English and in French, but to different constructions in just one language, namely German. So, the distance between meaning 1 and 2 is "1". Likewise, the distance between 1 and 3 is "2" because two of these languages treat them differently, and between 2 and 3 the distance is "1" because only French treats them differently. The establishment of the language-specific constructions and the counting of differences together are a metric on meanings, and the result is a list of distances between all pairs of meanings.

A different way of performing exactly the same calculation is obtained by a reformulation of the language-specific constructions into language-specific distance matrices. This reformulation might seem somewhat cumbersome at first, but it will allow for a much wider array of possible analyses—a few of which will be discussed in the next sections. The basic idea is to consider a language-specific construction to be a very simple notion of dissimilarity. As defined earlier, a construction can be considered to be a language-specific metric on expressions (cf. Table 1 and the discussion in Section 5). Such a metric only allows for the options “identical” (i.e. a dissimilarity/distance of “0”) or “different” (i.e. a dissimilarity/distance of “1”). From the perspective of English, the meanings 1, 2 and 3 are all identical (i.e. they belong to the same construction), which translates in a distance of zero between all pairs of these meanings. Of course, also the distance between each meaning and itself is zero (they necessarily belong to the same construction), so the result of reformulating the English first three meanings into a language-specific distance matrix is a matrix with all zeros (cf. the leftmost matrix in Figure 1—for convenience of presentation all matrices are shown completely, although distance matrices redundantly duplicate each entry in the upper and lower triangle). The same procedure can also be used for French and German, which will result in some distances of “1” because not all three meaning belong to the same class in these languages. Given these language-specific distance matrices, the cross-linguistic distance matrix on the meanings can now easily be computed by summing up these three matrices (cf. the rightmost matrix in Figure 1).[10]

Figure 1. language-specific constructions as distance matrices. Summing them together results in a cross-linguistic distance matrix on the meanings.

English

$$\begin{array}{cc} & \begin{matrix}1 & 2 & 3\end{matrix} \\ \begin{matrix}1\\2\\3\end{matrix} & \begin{bmatrix}0 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}\end{array}$$

French

$$\begin{array}{cc} & \begin{matrix}1 & 2 & 3\end{matrix} \\ \begin{matrix}1\\2\\3\end{matrix} & \begin{bmatrix}0 & 0 & 1\\0 & 0 & 1\\1 & 1 & 0\end{bmatrix}\end{array}$$

German

$$\begin{array}{cc} & \begin{matrix}1 & 2 & 3\end{matrix} \\ \begin{matrix}1\\2\\3\end{matrix} & \begin{bmatrix}0 & 1 & 1\\1 & 0 & 0\\1 & 0 & 0\end{bmatrix}\end{array}$$

Sum

$$\begin{array}{cc} & \begin{matrix}1 & 2 & 3\end{matrix} \\ \begin{matrix}1\\2\\3\end{matrix} & \begin{bmatrix}0 & 1 & 2\\1 & 0 & 1\\2 & 1 & 0\end{bmatrix}\end{array}$$

Doing these calculations for all 31 meanings in all 21 languages results in a 31×31 cross-linguistic distance matrix, giving the dissimilarity for all pairs of meanings—an ex-

---

[10] This reformulation opens up the possibility of comparing the structure of lexicalization between languages. This can be done by correlating the language-specific distance matrices from Figure 1. In effect, each distance matrix represents the language-specific perspective on the relation between the meanings. The similarity between two such matrices can be interpreted as a measure of how similarly languages deal with the coding of meanings. The details and implications of this approach to language comparison have to be left for another paper, though.

cerpt of which is shown in Table 4. The minimal value in this table is zero (i.e. the meanings belong to the same construction in all 21 languages), and the maximum is 21 (i.e. the meanings belong to different constructions in all 21 languages). These values can be normalized to the [0,1] interval by dividing them by 21 (shown in parentheses in the table). Just to give some perspective on these numbers, it appears that the pairs 'close'–'open', 'open'–'break' and 'close'–'break' are relatively similar (they belong to the same construction in about half of the languages investigated). In contrast, 'die/kill' is highly dissimilar from all others, as might have been expected, because the inchoative/causative alternation for this meaning is suppletive in most languages, and thus different from all other alternations in the same language.

Table 4. Excerpt of the cross-linguistic dissimilarity matrix on the meaning as established by summing over all 21 language-specific classifications.

	wake up	break	burn	die/kill	open	close
wake up	0	17 (.81)	16 (.76)	20 (.95)	17 (.81)	16 (.76)
break	17 (.81)	0	13 (.62)	19 (.90)	10 (.48)	12 (.57)
burn	16 (.76)	13 (.62)	0	20 (.95)	16 (.76)	17 (.81)
die/kill	20 (.95)	19 (.90)	20 (.95)	0	21 (1.0)	21 (1.0)
open	17 (.81)	10 (.48)	16 (.76)	21 (1.0)	0	10 (.48)
close	16 (.76)	12 (.57)	17 (.81)	21 (1.0)	10 (.48)	0

A complete analysis of the the full 31×31 distance matrix will not be pursued here, but one quick example will be given to indicate possible routes of analysis (see Cysouw 2008 for a more elaborate discussion). When multidimensional scaling is applied on the cross-linguistic distance matrix, then the first dimension (i.e. the dimension that explains most of the variation) appears to be related to the "scale of likelihood of spontaneous occurrence" (Haspelmath 1993: 105).[11] At one side of this scale predicates are found that prototypically do not need an agentive instigator, like 'boil', 'freeze', 'burn' (and in the multidimensional

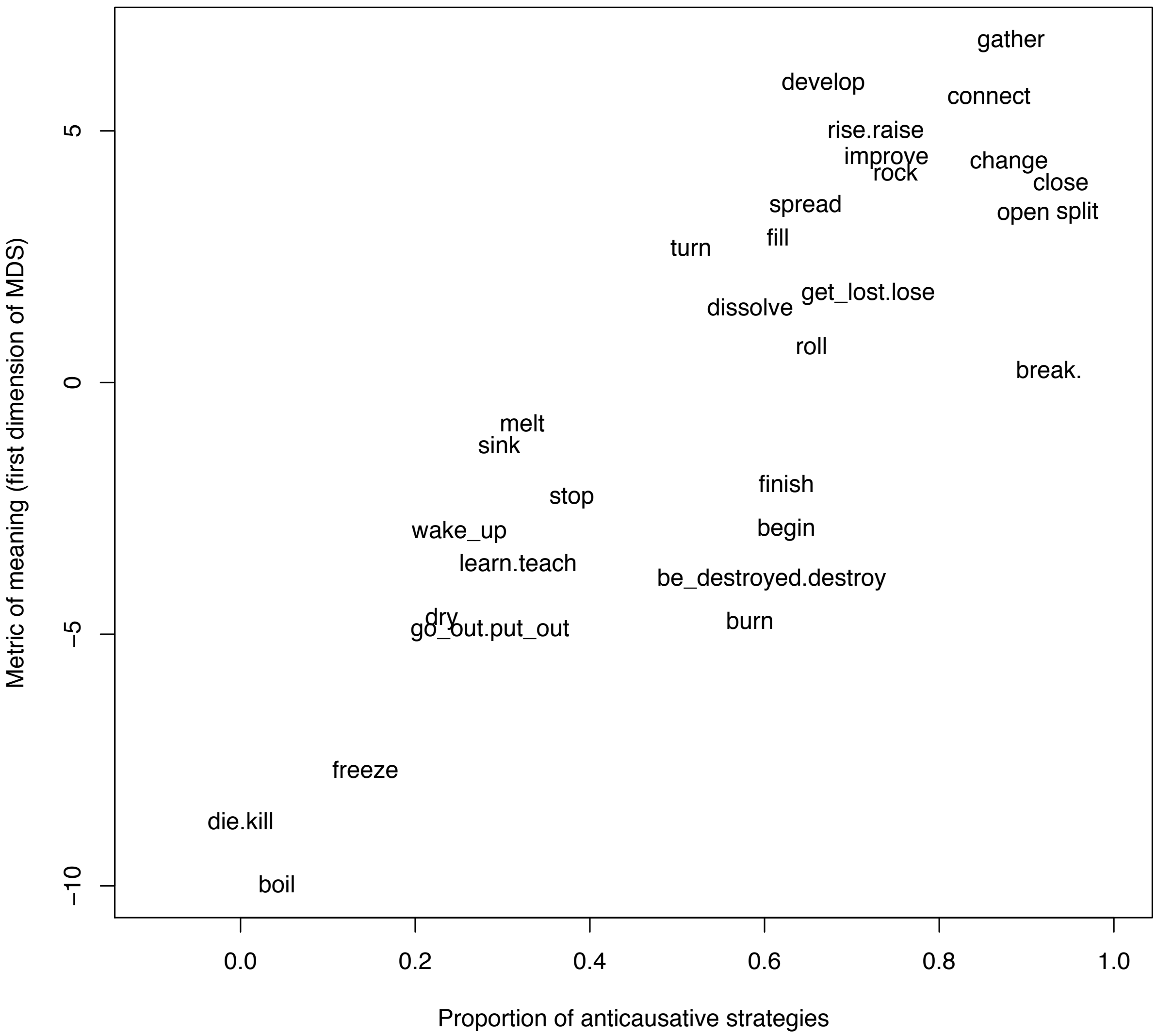
[11] For this calculation, classic multidimensional scaling was used through the implementation "cmdscale" in the statistical environment R (R Development Core Team 2007). All other calculations and graphs in this paper were also produced by using R.

scaling 'die/kill' is also found to belong to this side). The other side of the scale holds such events that normally have a human agent, like 'gather', 'connect' or 'change'. This scale was originally proposed by Haspelmath to explain the preference of certain meanings for particular behavioral strategies. Specifically, he argued that those meanings that are typically in need of a human instigator cross-linguistically have a preference for a causative coding strategy (i.e. the causative is derived from the inchoative), while the meanings at the other side of the scale have a preference for an anticausative strategy (i.e. the inchoative is derived from the causative).

Now, instead of deriving the scale of likelihood of spontaneous occurrence from behavioral strategies, as Haspelmath did, in this paper the scale is purely based on the analysis of language-specific constructions. The semantic scale of likelihood of spontaneous occurrence (here defined as the first dimension of the MDS of the metric on meaning) can then be correlated empirically with the proportion of languages that use an anticausative strategy (see Figure 2).[12] The correlation is almost perfect ($r = .83$, $p < 10^{-8}$). This example indicates that a linguistic scale can be conceived of as a (significant) correlation between meaning-similarity and form-similarity.

[12] Haspelmath, following up on earlier work by Nedjalkov, uses the fraction of anticausative by causative (A/C) strategies as an index for the cross linguistic preference for either of these strategies. The usage of this particular fraction is unfortunate because the resulting values are very unevenly distributed (they range between zero and infinite). I have used $A/(A+C)$ here instead. Another possibility would be to use $\log(A/C)$.

Figure 2. Correlation between preference for anticausative coding strategy and the first dimension of the MDS of the metric of meaning.



7.3. Metric B: Algorithmically approximating constructions

The reformulation of constructions as language-specific metrics on expressions, as discussed in relation to Figure 1 above, allows for a wide variety of other approaches to establishing a semantic map. The basic idea of this reformulation is that for each language a language-specific distance matrix is calculated, describing how similar the expressions of the meanings are from the perspective of each language individually. The cross-linguistic distances then are the result of simply summing up over all these language-specific distances. Using constructions, as done in the previous section, the language-specific matrices will only consist of “0” (indicating “same construction”) and “1” (indicating “different constructions”).

However, all values in between “0” and “1” can also be used, to indicate that two constructions are neither completely different nor completely similar. For example, one might argue that the German alternations *aufwachen/aufwecken* and *versinken/versenken* are different constructions, but also somewhat alike. They both involve a kind of ablaut, though the details are different. Neither considering them to be completely different, nor completely identical, will do justice to the empirical situation. To deal with such a situation, a gradient language-specific distance can be used. For example, one could set the language-specific distance between the two alternations above as 0.75 (see Table 5). The specification of gradient dissimilarities can be performed on the basis of a detailed analysis of each language individually. However, it is also possible to use a general method for measuring language-internal similarity. One such approach will be discussed in this section, and a simpler, but also less satisfying, method will be discussed in the next section.

Table 5. Different language-specific distances of some German inchoative/causative alternations.

		Yes/No distance				Gradient distance			
No.	German expressions	1	2	3	11	1	2	3	11
1	*aufwachen/aufwecken*	0	1	1	1	0	1	1	.75
2	*zerbrechen/zerbrechen*	1	0	0	1	1	0	0	1
3	*verbrennen/verbrennen*	1	0	0	1	1	0	0	1
11	*versinken/versenken*	1	1	1	0	.75	1	1	0

One possibility for comparing inchoative/causative alternations within the structure of a single language is to analyze each alternation as a collection of changes of letters needed to get from the inchoative to the causative string of letters. Changes are either a deletion of an existing letter or an insertion of a new letter. To match linguistic intuitions about what makes a similar change, the method distinguishes between making a change at the start of a word, at the end of a word, or in the middle of a word. For every inchoative/causative pair this leads to a list of changes how to get from the inchoative to the causative form. So, for example, to get from *rise* to *raise* only one change is needed, namely an <a> has to be inserted in the middle of the word. To compare two alternations, the number of shared letter changes is counted, and then normalized by the maximum number of changes attested. The distance between two alternations will then be the complement of this value (i.e. $1 - \text{shared}/\text{maximum}$).

For example, to get from the German inchoative *aufwachen* to causative *aufwecken* the following four changes are needed:

1) deletion of <a> inside the word ("*aufwchen*")
2) deletion of <h> inside the word ("*aufwcen*")
3) insertion of <e> inside the word ("*aufwecen*")
4) insertion of <k> inside the word ("*aufwecken*")

To get from from German inchoative *versinken* to causative *versenken* the following two changes are needed:

1) deletion of <i> inside the word ("*versnken*")
2) insertion of <e> inside the word ("*versenken*")

These two sets of changes have one change in common ("insertion of <e> inside the word"), and the maximum number of changes needed is "4" (for the *aufwachen/aufwecken* alternation), so the distance between the two alternations is $1 - 1/4 = .75$ (cf. Table 5). This algorithm could be improved in various ways.[13] However, the main point is that it is relatively easy to get a rough estimate of the language-internal dissimilarity between two inchoative/causative alternations.[14]

To get from language-specific dissimilarities to a cross-linguistic distance matrix, all individual matrices are summed together. An excerpt of the resulting matrix is shown in Table 6, which can be compared with the same selection shown in Table 4. Although the two tables are not completely identical, the values are astonishingly close. The complete correla-

---

[13] There are various questionable decisions being made in this algorithm. First, it operates on letters, where ideally it would work on sounds. Second, there is no reason to restrict the algorithm to only insertions and deletions—also exchanges could be used, or other operations. Further, every insertion and deletion is equally weighted, though some might be more significant than others. And instead of dividing by the maximum number of changes one could also use another normalization, like dividing by the average number of changes.

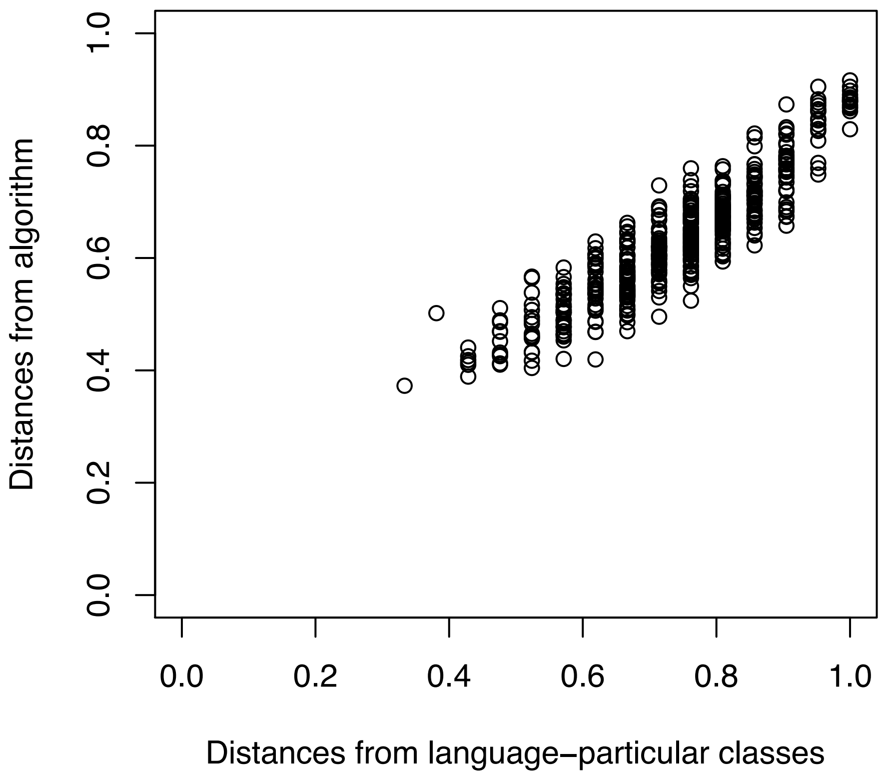
[14] I thank Hagen Jung for assistance with the implementation of this algorithm.

tion between the results of this algorithmic notion of dissimilarity and the dissimilarity based on the manually established language-specific constructions is shown in Figure 3 ($r = .91$). Shown on the x-axis in this figure are the dissimilarities ("distances") from the metric discussion in the previous Section 7.2. On the y-axis, the distances from the algorithmic approach as discussed in this section are shown. The close match between these two methods suggests that automatic approaches can be very useful in the establishment of cross-linguistic metrics on meaning. In general, it appears that the errors introduced by the linguistically naive algorithm are easily corrected by summing up over many languages.

Table 6. Excerpt of the cross-linguistic distance matrix as established by the algorithmic approach.

	wake up	break	burn	die/kill	open	close
wake up	0	14.1 (.67)	14.5 (.69)	18.5 (.88)	13.8 (.66)	13.5 (.64)
break	14.1 (.67)	0	12.7 (.61)	17.5 (.83)	10.2 (.49)	10.8 (.51)
burn	14.5 (.69)	12.7 (.61)	0	17 (.81)	14.5 (.69)	15.4 (.73)
die/kill	18.5 (.88)	17.5 (.83)	17 (.81)	0	18.7 (.89)	18.6 (.89)
open	13.8 (.66)	10.2 (.49)	14.5 (.69)	18.7 (.89)	0	10.3 (.49)
close	13.5 (.64)	10.8 (.51)	15.4 (.73)	18.6 (.89)	10.3 (.49)	0

Figure 3. Correlation between cross-linguistic distances as establish by language-specific classes and by the algorithmic approach.

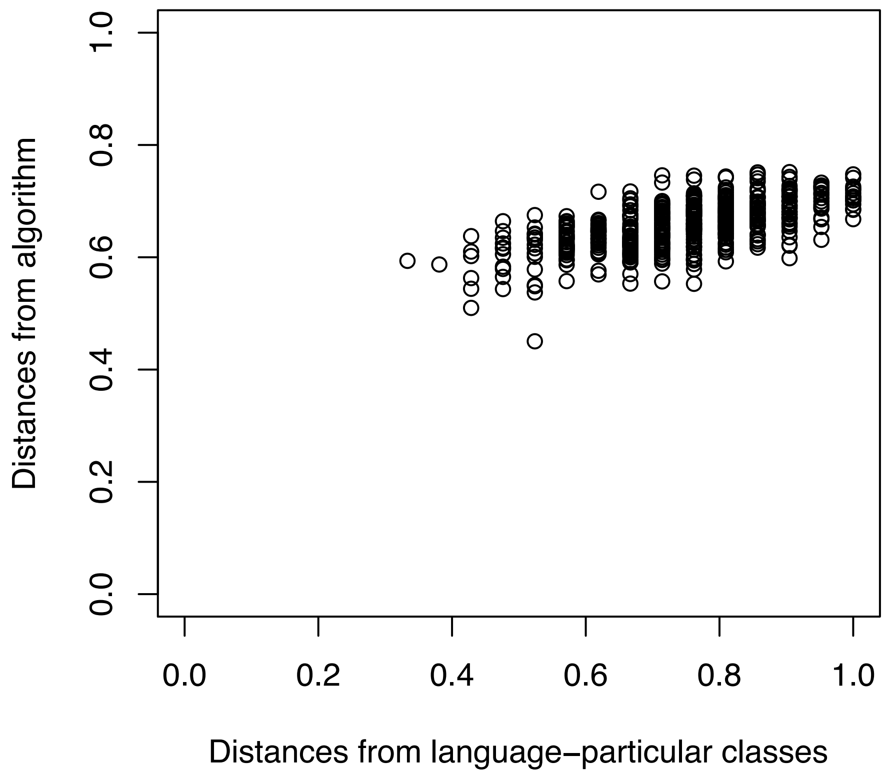


7.4. Metric C: Simplistic string-based similarity

The good results of the algorithmic approach to establishing language-specific similarities prompted me to try out an even simpler, even more linguistically naive algorithmic approach. It is based on the LONGEST COMMON SUBSTRING measure of similarity between two strings of letters. This similarity consists of the length of the longest consecutive stretch of letters shared between to expression. So, for example, *house* and *mouse* share 4 letters in a row. To use this measure of similarity for inchoative/causative alternations, I pasted the inchoative and the causative forms together into one string without spaces (e.g. French *seréveillerréveiller* or *sebriserbriser*) and established the longest common substring (in the French example this would be "2" for the string "*se*"). This approach of course finds all kinds of small random similarities (e.g. *wakeupwakeup* and *breakbreak* also have a longest common substring of "2" for the string "*ak*") and in general it only works well with concatenative morphology or morphologically independent markers (like the reflexive *se* in the French example above).

Figure 4 shown the relation between the distances from this very simplistic approach (shown on the y-axis) to the distances from the linguistically sophisticated approach using language-specific classes, as discussed in Section 7.2. The match between this extremely simple measurement of language-specific similarity to the linguistically sophisticated similarity using language-specific classes is not as good as for the more elaborate algorithmic approach from the previous section ($r = .61$, cf. Figure 4 with the previous Figure 3), though the correlation is still highly significant (Mantel test $p < .00001$), indicating that even with linguistically very naive similarity measures relatively good overall results are possible.

Figure 4. Correlation between cross-linguistic distances as establish by language-specific classes and by the longest common substring.



8. Conclusion

By using the world's linguistic diversity, the study of meaning can be transformed from an introspective inquiry into an subject of empirical investigation. For this to be possible, the notion of meaning has to be operationalised by defining the meaning of an expression as the collection of all contexts in which the expression can be used. Under this definition, meaning can be empirically investigated by sampling contexts. A semantic map is a technique to show the relations between such sampled contexts. Or, formulated more technically, a semantic map is a visualization of a metric on contexts sampled to represent a domain of meaning. Or, put more succinctly, a semantic map is a metric on meaning.

The relation between different contexts/meanings can be investigated by looking at their expressions in many languages. The more similar these expressions, when averaged over all languages studied, the more similar the contexts. So, to investigate the similarity between contexts, only judgments about the local similarity between expressions within the structure of individual languages are needed. In general, this similarity between language-specific expressions is a special—language-specific—metric between contexts. A metric on meaning, then, is the cross-linguistic average of many language-specific expression metrics.

A language-specific expression metric can be very fine-grained, and to a large extent automatically retrieved, opening up the possibility to speed up the empirical study of meaning. It is important to realize, however, that for any resulting semi-automatically retrieved metric on meaning, the interpretation ("the meaning of the metric") is of course still in the eye of the beholder, namely, the human investigator.

Appendix: Language-specific classes of causative/inchoative alternations

Arabic

Class A: C/CC
1. saḧaa/saḧḧaa
8. darasa/darrasa
14. damara/dammara
31. waqafa/waqqafa

Class B: in/ø
2. inkasara/kasara
5. infataḧa/fataḧa
6. inqafala/qafala
13. inṣahara/ṣahara
30. inšaqqa/šaqqa

Class C: in/ʔ
3. iḧtaraqa/ʔaḧraqa
20. inṭafaʔa/ʔaṭfaʔa
22. intahaa/ʔanhaa

Class D: t/ø
9. iltamma/lamma
10. intašara/našara
17. irtabaṭa/rabaṭa
21. irtafaʕa/rafaʕa
27. imtalaʔa/malaʔa

Class E: ø/ʔ
11. ġariqa/ʔaġraqa
18. ġalaa/ʔaġlaa
23. daara/ʔadaara
26. ðaaba/ʔaðaaba

Class F: ta/ø
12. tabaddala/baddala
16. taṭawwara/ṭawwara
19. taʔarjaḧa/ʔarjaḧa
24. tadaḧraja/daḧraja
25. tajammada/jammada
28. taḧassana/ḧassana

Singular classes:
4. maata/qatala
7. badaʔa
15. daaʕa/xasira
29. jaffa/jaffafa

Armenian

Class A: ø/c
1. artnanal/artnacnel
16. zarzanal/zarzacnel
21. barʒranal/barʒracnel
22. k'eršanal/k'eršacnel
28. lavanal/lavacnel
29. čoranal/čoracnel

Class B: v/ø
2. ǯardvel/ǯardel
3. ayrvel/ayrel
6. pak'vel/pak'el
7. sksvel/sksel
9. havakvel/havakel
10. əndarc'ak'vel/əndarc'ak'el
11. xegolvel/xegolel
12. poxvel/poxel
13. halvel/halel
14. kandvel/kandel
17. k'ap'vel/k'ap'el
19. č'oč'vel/č'oč'el
23. pttvel/pttel
24. glorvel/glorel
26. luc'vel/luc'el
30. č'eɣkvel/č'eɣkel

Class C: v/n
5. bacvel/bacanal
27. lcvel/lcnel

Class D: ø/Vcn
8. sovorel/sovorecnel
18. eṙal/eṙacnel
31. k'angnil/k'angnecnel

Class E: č/cn
15. k'orčel/k'orcnel
20. hangčel/hangcnel
25. saṙčel/saṙecnel

Class F:
4. spanel/mernel

English

Class A: Identical
1. wake up
2. break
3. burn
5. open
6. close
7. begin
9. gather
10. spread
11. sink
12. change
13. melt
16. develop
17. connect
18. boil
19. rock
22. finish
23. turn
24. roll
25. freeze
26. dissolve
27. fill
28. improve
29. dry
30. split
31. stop

Singular classes:
4. die/kill
8. learn/teach
14. be destroyed/destroy
15. get lost/lose
20. go out/put out
21. rise/raise

Finnish

Class A: ø/tt
1. herätä/herättää
3. palaa/polttaa
8. oppia/opettaa
10. levitä/levittää
13. sulaa/sulattaa
18. kiehua/kiehuttaa
19. kiikkua/kiikuttaa
20. sammua/sammuttaa
21. kohota/kohottaa
22. loppua/lopettaa
24. vieriä/vierittää
26. liueta/liuottaa
29. kuivaa/kuivata

Class B: U/ø
2. murtua/murtaa
12. muuttua/muuttaa
16. kehittyä/kehittää
23. vääntyä/vääntää
27. täyttyä/täyttää
28. parantua/parantaa

Class C: UtU/ø
5. avautua/avata
6. sulkeutua/sulkea
14. tuhoutua/tuhota

Class D: ntu/t
9. kokoontua/koota
15. hukkaantua/hukata

Class E: tyä/dytää
17. yhtyä/yhdistää
25. jäätyä/jäädyttää
31. pysähtyä/pysähdyttää

Singular classes:
4. kuolla/tappaa
7. alkaa/aloitaa
11. laskea
30. haljeta/halkaista

French

Class A: se/ø
1. se réveiller/réveiller
2. se briser/briser
5. s'ouvrir/ouvrir
6. se fermer/fermer
9. s'assembler/assembler
10. s'étendre/étendre
11. s'enfoncer/enfoncer
15. se perdre/perdre
16. se développer/développer
17. se lier/lier
19. se balancer/balancer
20. s'éteindre/éteindre
21. se lever/lever
23. se tourner/tourner
26. se dissoudre/dissoudre
27. se remplir/remplir
28. s'améliorer/améliorer
30. se fendre/fendre
31. s'arrêter/arrêter

Class B: Identical
3. brûler
7. commencer
8. apprendre
12. changer
22. finir
24. rouler
25. geler
29. sécher

Class C: ø/faire
13. fondre/faire fondre
18. bouillir/faire bouillir

Singular classes:
4. mourir/tuer
14. être détruit/détruir

Georgian

Class A: i/a
1. gaiɣviʒebs/gaaɣviʒebs
8. isc'avlis/asc'avlis

Class B: i + a/a + s
2. imt'vreva/amt'vrevs
5. gaiɣeba/gaaɣebs
11. daixrčoba/axrčobs
14. daingreva/daangrevs
19. irxeva/arxevs
27. aivseba/aavsebs
30. gaip'oba/gaap'obs

Class C: i + eba/ø + avs
6. daixureba/daxuravs
15. ik'argeba/k'argavs
25. gaiqineba/gaqinavs

Class D: i + eba/ø + is
9. šeik'ribeba/šek'rebs
12. šeicvleba/šecvlis
16. daišleba/dašlis
26. gaixsneba/gaxsnis

Class E: ø + eba/a + obs
13. gadneba/gaadnobs
20. kreba/akrobs
29. šreba/ašrobs

Class F: ø + deba/a + ebs
10. gavrceldeba/gaavrcelebs
22. gatavdeba/gaatavebs
28. gaumǯobesdeba/ gaaumǯobesebs
31. gačerdeba/gaačerebs

Class G: ø + avs/a + ebs
23. brunavs/abrunebs
24. migoravs/miagorebs

Singular classes:
3. ic'vis/c'vavs
4. mok'vdeba/mok'lavs
7. daic'qeba/daic'qebs
17. šeexameba/šeuxamebs
18. duɣs/aduɣebs
21. adgeba/aiɣebs

German

Class A: Identical
2. zerbrechen
3. verbrennen
7. anfangen
13. schmelzen
18. kochen
19. schaukeln
24. rollen
25. einfrieren
29. trocknen
31. anhalten

Class B: sich/ø
5. sich öffnen/öffnen
6. sich schliessen/schliessen
9. sich sammeln/sammeln
10. sich ausbreiten/ausbreiten
12. sich verändern/verändern
16. sich entwickeln/entwickeln
17. sich verbinden/verbinden
21. sich heben/heben
23. sich umdrehen/umdrehen
26. sich auflösen/auflösen
27. sich füllen/füllen
28. sich verbessern/verbessern
30. sich spalten/spalten

Singular classes:
1. aufwachen/aufwecken
4. sterben/töten
8. lernen/lehren
11. versinken/versenken
14. kaputt gehen/ kaputt machen
15. verloren gehen/verlieren
20. erlöschen/löschen
22. enden/beenden

Greek

Class A: Identical
1. ksipnó
2. spázo
5. anígho
6. klíno
7. arçízo
8. mathéno
12. alázo
14. xalnó
18. vrázo
20. svíno
22. telióno
23. yirízo
25. paghóno
27. yemízo
30. xorízo
31. stamatáo

Class B: me/ø
3. kéome/kéo
9. singendrónome/ singendróno
10. dhiadhídhome/dhiadhídho
11. vithízome/vithízo
13. tíkome/tíko
15. xánome/xáno
16. anaptísome/anaptíso
17. sindhéome/sindhéo
19. liknízome/liknízo
21. sikónome/sikóno
24. kiliéme/kilió
26. dhialíome/dhialío
28. veltiónome/veltióno
29. apoksirénome/apoksiréno

Singular classes:
4. pethéno/skotóno

Hebrew

Class A: hit/ø
1. hitʕorer/ʕorer
9. hitʔasef/ʔasaf
10. hitpares/paras
12. hištana/šina
16. hitpatah́/patah́
17. hitkašer/kišer
19. hitnadned/nidned
21. hitromem/romem
23. histovev/sovev
26. hitporer/porer
27. hitmale/mile
28. hištaper/šiper
29. hityabeš/yibeš
30. hitpacel/picel

Class B: ni/ø
2. nišbar/šavar
3. nisraf/saraf
5. niftah́/patah́
6. nisgar/sagar
22. nigmar/gamar
31. neʕecar/ʕacar

Class C: ø/hV
4. mat/hemit
14. h́arav/heh́eriv
18. ratah́/hirtiah́
25. kafa/hikfi

Class D: av/ib
11. tavaʕ/tibaʕ
15. ʔavad/ʔibed
20. kava/kiba

Singular classes:
7. hith́il
8. lamad/limed
13. namas/hemes
24. nagol/galal

Hindi-Urdu

Class A: ø/aa
1. jaagnaa/jagaanaa
3. jalnaa/jalaanaa
8. parhnaa/parhaanaa
10. phailnaa/phailaanaa
13. pighalnaa/pighlaanaa
19. hilnaa/hilaanaa
21. uṭhnaa/uṭhaanaa
23. phirnaa/phiraanaa
24. luṛhaknaa/luṛhkaanaa
25. jamnaa/jamaanaa
26. ghulnaa/ghulaanaa
29. suukhnaa/sukhaanaa

Class B: ṭ/ṛ
2. ṭuuṭnaa/ṭoṛnaa
30. phaṭnaa/phaaṛnaa

Class C: a/aa
4. marnaa/maarnaa
14. ujarnaa/ujaarnaa
17. bandhnaa/baandhnaa
18. ubalnaa/ubaalnaa

Class D: u/o
5. khulnaa/kholnaa
31. ruknaa/roknaa

Class E: honaa/karnaa
6. band honaa/band karnaa
7. šuruu honaa/
   šuruu karnaa
9. ikaṭṭhaa honaa/
   ikaṭṭhaa karnaa
16. vikaas honaa/
   vikaas karnaa
20. gul honaa/gul karnaa
22. xatm honaa/xatm karnaa
28. behtar honaa/
   behtar banaanaa

Class F: Identical
12. badalnaa
27. bharnaa

Singular classes:
11. ḍuubnaa/ḍubonaa
15. khojaanaa/khonaa

Hungarian

Class A: d/szt
1. felébred/felébreszt
10. terjed/terjeszt
11. elsüllyed/elsüllyeszt
13. olvad/olvaszt

Class B: ø/Vt
3. elég/eléget
15. elvész/elveszít
23. forog/forgat
31. megáll/megállít

Class C: Vlik/it
5. kinyílik/kinyit
9. összegyülik/összegyüjt

Class D: Odik/ø
6. záródik/zár
7. elkezdödik/elkezd
22. befejezödik/befejez
26. oldódik/old

Class E: Ul/it
8. tanul/tanít
14. elpusztul/elpusztít
24. gurul/gurít
28. javul/javít

Class F: ik/tat
12. megváltozik/megváltoztat
19. hintázik/hintáztat;

Class G: ad/it
29. szárad/szárít
30. széthasad/széthasít

Singular classes:
2. összetörik/összetör
4. meghal/megöl
16. fejlödik/fejleszt
17. szövetkezik/összeköt
18. fö/föz
20. kialszik/kiolt
21. emelkedik/emel
25. megfagy/megfagyaszt
27. megtelik/tölt

Indonesian

Class A: ter/me + kan
1. terbangun/membangunkan
10. tersebar/menyebarkan

Class B: ø/me + kan
2. patah/mematahkan
4. mati/mematikan
11. tenggelam/
   menenggelamkan
14. binasa/membinasakan
20. padam/memadamkan
22. selesai/menyelesaikan
26. larut/melarutkan
29. kering/mengeringkan

Class C: ter/me
3. terbakar/membakar
5. terbuka/membuka
27. terisi/mengisi
30. terbelah/membelah

Class D: ø/me
6. tutup/menutup
7. mulai/memulai

Class E: ber/meng
8. belajar/mengajar
12. berubah/mengubah
19. berayun/mengayun

Class F: ø/kan
9. mengumpul/mengumpulkan
13. mencair/mencairkan
24. menggelinding/
   menggelindingkan
25. membeku/membekukan

Class G: ber/me + kan
16. berkembang/
   mengembangkan
17. bergabung/menggabungkan
23. berbalik/membalikkan
31. berhenti/menghentikan

Singular classes:
15. menghilang/kehilangan
18. direbus/merebus
21. kenaikan/menaikkan
28. bertambahbaik/
   memperbaiki

Japanese

Class A: Vr/Vs
1. okiru/okosu
6. toziru/tozasu
13. tokeru/tokasu
19. yureru/yurasu
20. kieru/kesu
23. mawaru/mawasu
24. korogaru/korogasu
26. tokeru/tokasu
27. mitiru/mitasu
28. naoru/naosu

Class B: er/ø
2. oreru/oru
3. yakeru/yaku
30. sakeru/saku

Class C: ø/er
5. aku/akeru
11. sizumu/sizumeru

Class D: a/e
7. hazimaru/hazimeru
8. osowaru/osieru
9. atumaru/atumeru
10. hirogaru/hirogeru
12. kawaru/kaeru
17. tunagaru/tunageru
21. agaru/ageru
22. owaru/oeru
31. tomaru/tomeru

Class E: ø/ase
16. hattatu suru/ hattatu saseru
25. kooru/kooraseru

Class F: ø/as
18. waku/wakasu
29. kawaku/kawakasu

Singular classes:
4. sinu/korosu
14. kowareru/kowasu
15. nakunaru/nakusu

Lezgian

Class A: Identical
2. xun
3. kun
4. q’in
18. rugun
30. xun

Class B: x̂/ø
5. aq^{h}a x̂un/aq^{h}ajun
6. k’ew x̂un/k’ewun
7. bašlamiš x̂un/bašlamišun
8. čir x̂un/čirun
9. k’wat’ x̂un/k’wat’un
19. e’čä x̂un/e’čäğun
21. xkaž x̂un/xkažun
22. kütäh x̂un/kütähun

Class C: ø/r
10. čuk’un/čuk’urun
13. c’urun/c’ururun
14. čuk’un/čuk’urun
17. sadsadaw q’un/sadsadaw q’urun
20. tüxün/tüxürun
23. elqün/elqürun
25. č’agun/č’agurun
26. c’urun/c’ururun
27. ac’un/ac’urun
29. q’urun/q’ururun

Class D: x̂/ar
11. batmiš x̂un/batmišarun
12. degiš x̂un/degišarun
28. q^{h}san x̂un/q^{h}sanarun

Class E: ø/ar
15. kwax̂un/kwadarun
31. aqwazun/aqwazarun

Class F: fin/raqurun
16. wilik fin/wilik raqurun
24. awax̂izawax̂iz fin/awax̂iza-wax̂iz raqurun

Class D: t/d
1. axwaraj awatun/ axwaraj awudun

Lithuanian

Class A: ø/in
1. pabusti/pabudinti
3. degti/deginti
11. skendeti/skandinti
18. virti/virinti
20. gesti/gesinti
26. ištirpti/ištirpinti
28. gerėti/gerinti
29. sausti/sausinti

Class B: ūp/au
2. lūžti/laužti
14. sugriūti/sugriauti
31. nutrūkti/nutraukti

Class C: si/ø
5. atsidaryti/atidaryti
7. prasidėti/pradėti
10. išsiplėsti/išplėsti
12. pasikeisti/pakeisti
13. išsilydyti/išlydyti
15. pasimesti/pamesti
22. pasibaigti/pabaigti
27. prisipildyti/pripildyti

Class D: s/ø
6. klostytis/klostyti
8. mokytis/mokyti
9. rinktis/rinkti
16. plėtotis/plėtoti
17. jungtis/jungti
19. suptis/supti
23. suktis/sukti
24. ristis/risti

Class E: i/e
21. pakilti/pakelti
30. perskilti/perskelti

Singular classes:
4. užmušti/mirti
25. užšalti/užšaldyti

Mongolian

Class A: ø/V
1. serex/sereex
3. šatax/šataax
20. untrax/untraax
25. xöldöx/xöldööx
29. xatax/xataax
31. zogsox/zogsoox

Class B: r/l
2. xugarax/xugalax
30. xagarax/xagalax

Class C: Vgd/ø
6. xaagdax/xaax
12. öörčlögdöx/öörčlöx
15. xajagdax/xajax
17. xolbogdox/xolbox
21. örgögdöx/örgöx

Class D: ø/g
7. üüsex/üüsgex
8. surax/surgax
18. buclax/bucalgax
22. duusax/duusgax
26. uusax/uusgax
27. düürex/düürgex

Class E: ø/UUl
9. cuglax/cugluulax
11. živex/živuulex
13. xajlax/xajluulax
16. xögžix/xögžüülex
19. dajvalzax/dajvalzuulax
23. ergex/ergüülex
24. önxröx/önxrüülex
28. sajžrax/sajžruulax

Class F: r/ø
10. delgerex/delgex
14. evdrex/evdex

Singular classes:
4. üxex/alax
5. ongojx/ongojlgox

Rumanian

Class A: se/ø
1. se trezi/trezi
2. se rupe/rupe
5. se deschide/deschide
6. se închide/închide
9. se aduna/aduna
10. se răspîndi/răspîndi
11. se scufunda/scufunda
12. se schimba/schimba
13. se topi/topi
15. se pierde/pierde
16. se dezvolta/dezvolta
17. se uni/uni
19. se legăna/legăna
20. se stinge/stinge
21. se ridica/ridica
22. se sfîrşi/sfîrşi
23. se învîrti/învîrti
24. se rostogoli/rostogoli
26. se dizolva/dizolva
27. se umple/umple
28. se îndrepta/îndrepta
29. se usca/usca
30. se crăpa/crăpa
31. se opri/opri

Class B: Identical
3. arde
7. începe
18. fierne

Singular classes:
4. muri/ucide
8. învăţa/preda
14. ?/distruge
25. îngheţa/face sa îngheţe

Russian

Class A: sja/ø
2. lomat'sja/lomat'
5. otkryt'sja/otkryt'
6. zakryt'sja/zakryt'
7. načat'sja/načat'
8. učit'sja/učit'
9. sobrat'sja/sobrat'
10. rasprostranit'sja/
    rasprostranit'
12. izmenit'sja/izmenit'
13. rasplavit'sja/rasplavit'
14. razručit'sja/razručit'
15. terjat'sja/terjat'
16. razvit'sja/razvit'
17. sočetat'sja/sočetat'
19. kačat'sja/kačat'
21. podnjat'sja/podnjat'
22. končit'sja/končit'
23. povernut'sja/povernut'
24. katit'sja/katit'
26. rastvorit'sja/rastvorit'
27. napolnit'sja/napolnit'
28. ulučšit'sja/ulučšit'
30. raskolot'sja/raskolot'
31. ostanovit'sja/ostanovit'

Class B: nut/it
11. utonut'/utopit'
20. gasnut'/gasit'
25. zamerznut'/zamorozit'
29. soxnut'/sušit'

Singular classes:
1. prosnut'sja/budit'
3. goret'/žeč'
4. umeret'/ubit'
18. kipet'/kipjatit'

Swahili

Class A: k/sh
1. amka/amsha
13. yeyuka/yeyusha
18. chemka/chemsha
24. fingirika/fingirisha
26. yeyuka/yeyusha
29. kauka/kausha

Class B: k/ø
2. vunjika/vunja
3. unguka/ungua
5. funguka/fungua
9. kusanyika/kusanya
14. haribika/haribu
20. zimika/zima
21. inuka/inua
22. malizika/maliza
12. geuka/geua
30. pasuka/pasua

Class C: w/ø
6. fungwa/funga
17. ungwa/unga

Class D: ø/sh
7. anza/anzisha
8. funda/fundisha
11. zama/zamisha
16. sitawia/sitawisha
19. yonga/yongesha
23. zungua/zungusha
25. ganda/gandisha
31. simama/simamisha

Class E: ø/z
10. enea/eneza
15. potea/poteza
27. jaa/jaza

Singular classes:
4. fa/ua
28. fanya ujambo/ pata ujambo

Turkish

Class A: ø/dVr
1. uyanmak/uyandırmak
4. ölmek/öldürmek
20. sönmek/söndürmek
21. kalkmak/kaldırmak
23. dönmek/döndürmek
25. donmak/dondurmak
27. dolmak/doldurmak
31. durmak/durdurmak

Class B: Vl/ø
2. kırılmak/kırmak
5. açılmak/açmak
10. yayılmak/yaymak
14. bozulmak/bozmak
26. çözülmek/çözmek

Class C: n/ø
9. toplanmak/toplamak
19. sallanmak/sallamak
24. yuvarlanmak/yuvarlamak

Class D: n/t
6. kapanmak/kapatmak
8. öğrenmek/öğretmek

Class E: ø/tir
16. inkişaf etmek/inkişaf ettirmek
12. degişmek/degiştirmek
17. birleşmek/birleştirmek

Class F: ø/ir
11. batmak/batırmak
18. pişmek/pişirmek
22. bitmek/bitirmek

Class G: ø/t
13. erimek/eritmek
28. düzelmek/düzeltmek
29. kurumak/kurutmak
30. çatlamak/çatlatmak

Singular classes:
3. yanmak/yakmak
7. ?/başlamak
15. kaybolmak/kaybetmek

Udmurt

Class A: ø/ty
1. sajkany/sajkatyny
8. dyšyny/dyšetyny
10. võlmyny/võlmytyny
11. vyjyny/vyjytyny
13. ćyžany/ćyžatyny
14. kuaškany/kuaškatyny
15. ysyny/ystyny
23. bergany/bergatyny
26. sylmyny/sylmytyny
27. tyrmyny/tyrmytyny
31. dugdyny/dugdytyny

Class B: śky/ø
2. tijaśkyny/tijany
3. sutskyny/sutyny
5. ustiśkyny/ustyny
6. pytsaśkyny/pytsany
9. l'ukaśkyny/l'ukany
12. voštiśkyny/voštyny
17. gerʒaskyny/gerʒany
19. vettaśkyny/vettany
21. ʒutśkyny/ʒutyny
30. pil'iśkyny/pil'yny

Class C: Identical
7. kutskyny
18. byrektyny
20. kysyny

Class D: sky/ty
16. azinskyny/azintyny
24. pityrskyny/pityrtyny
28. umojatskyny/umojatyny

Class E: my/ty
22. bydesmyny/bydestyny
25. kynmyny/kyntyny
29. kuasmyny/kuastyny

Class F:
4. kulyny/viyny

References

Comrie, Bernard & Tania Kuteva. 2005. Relativization strategies, in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures.* Oxford: Oxford University Press, 398-405.

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

———. 2003. *Typology and Universals.* (second edition). (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

———. 2007. Exemplar semantics. Unpublished manuscript. available online at http://www.unm.edu/~wcroft/WACpubs.html.

Croft, William & Keith T. Poole. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis, *Theoretical Linguistics* 34(1): 1-37.

Cysouw, Michael. 2007. Building semantic maps: the case of person marking, in: Bernhard Wälchli & Matti Miestamo (eds.) *New Challenges in Typology.* (Trends in Linguistics: Studies and Monographs, 189). Berlin: Mouton de Gruyter, 225-248.

———. 2008. Generalizing scales, in: Marc Richards & Andrej Malchukov (eds.) *Scales.* (Linguistische Arbeits Berichte, 86). Leipzig: Institut für Linguistik, Universität Leipzig, 379-396.

Dahl, Östen. 1985. *Tense and Aspect systems.* Oxford: Blackwell.

Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language.* Oxford: Oxford University Press.

Haiman, John. 1985. *Natural Syntax.* Cambridge: Cambridge University Press.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations, in: Bernard Comrie & Maria Polinsky (eds.) *Causatives and Transitivity.* (Studies in Language Companion Series, Amsterdam: Benjamins, 87-120.

———. 1997. *Indefinite Pronouns.* (Oxford Studies in Typology and Linguistic Theory). Oxford: Clarendon.

———. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison, in: Michael Tomasello (eds.) *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure (Volume 2).* Mahwah, NJ: Erlbaum, 211-242.

———. 2008. Comparative concepts and descriptive categories in cross-linguistic studies. Unpublished manuscript. available online at http://www.eva.mpg.de/~haspelmt/papers.html.

Keenan, Edward L. 1976. Towards a universal definition of ‘subject’, in: Charles N. Li (eds.) *Subject and Topic*. New York, NY: Academic Press, 303-333.

Keenan, Edward L. & Bernard Comrie. 1977. Noun phrase accessibility and universal grammar, *Linguistic Inquiry* 8(1): 63–99.

Koptjevskaja-Tamm, Maria. 1993. *Nominalizations*. London: Routledge.

———. 2005. Action nominal constructions, in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures*. Oxford: Oxford University Press, 254-257.

Lehmann, Christian. 1984. *Der Relativsatz: Typologie seiner Strukturen; Theorie seiner Funktionen; Kompendium seiner Grammatik*. Tübingen: Narr.

Levinson, Stephen C. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. (Language, Culture & Cognition, 5). Cambridge: Cambridge University Press.

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

Majid, Asifa, Melissa Bowerman, Miriam van Staden, & James S. Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective, *Cognitive Linguistics* 18(2): 133-152.

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Wälchli, Bernhard. 2005. *Co-compounds and Natural Coordination*. Oxford: Oxford University Press.

Wälchli, Bernhard & Michael Cysouw. 2008. Toward a semantic map of motion verbs. Unpublished manuscript. available online at http://www.eva.mpg.de/~cysouw/publications.html.

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals*. Oxford: Oxford University Press.