

Generalizing Scales

Michael Cysouw

Max Planck Institute for Evolutionary Anthropology

Abstract

Instead of considering scales to be linearly ordered structures, it is proposed that scales are better conceived of as metrics (dissimilarity matrices). Further, to be considered a scale of typological interest, there should be a significant correlation between a meaning-scale and a form-scale. This conceptualisation allows for a fruitful generalization of the concept "scale". As a hands-on example of the proposals put forward in this paper, the "scale of likelihood of spontaneous occurrence" (Haspelmath 1993) is reanalyzed. This scale describes the prototypical agentivity of the subject of a predicate.

1. Scales as restrictions on form-function mapping

Scales¹ of linguistic structure are one of the more promising avenues of research into the unification of the worldwide linguistic diversity. Although our growing understanding of the diversity of the world's languages seems to put more and more doubt on many grandiose attempts on universally valid generalizations, the significance of scales for human languages (like the well-known animacy scale) still appears to stand strong. So, what actually is a scale? A scale seems to be mostly thought of as an asymmetrical one-dimensional arrangement (a "total order" in mathematical parlance) on certain cross-linguistic categories/functions. Put differently, a scale is a linear ordering of functions with a "high end" and a "low end". To be considered an interesting scale, the formal encoding of these functions in actual languages should be related to this linear ordering.

In this paper, I will argue that this concept of a scale can be fruitfully generalized. In a very general sense, all linguistic structure consists of forms expressing particular functions. If we find restrictions—across languages—on the kind of forms that are used to express certain functions, then this

¹The term "scale" is used here synonymously to what is also known as an "implicational hierarchy", "markedness hierarchy" or simply "hierarchy" in linguistics.

amounts to a cross-linguistic generalization. I would like to suggest that every restriction on the form-function mapping can be considered to be a (generalized) scale. Traditional one-dimensional scales are just a special kind of such cross-linguistic restrictions on form-function mapping. When the limitation that a scale has to be one-dimensional is discarded, and the concept of ordering is exchanged for a concept of distance, then the notion of a scale can be nicely generalized to cover many, if not all, restrictions on form-function mapping (cf. Croft, 2003, 133-142).

To establish a scale in its generalized conception, it is necessary, first, to establish a cross-linguistic scale of functions; second, to establish a cross-linguistic scale of forms; and, third, to observe a match between the two. Strictly speaking, a cross-linguistic scale, or hierarchy, is the interpretation of any such observed match. These three topics—scales of function, scales of form, and matching them—will be discussed in turn below in Section 4 to Section 6. However, first I will introduce a central tool for the generalization of scales, namely the dissimilarity matrix (Section 2), and the concrete example to be used for the discussion of the generalization, namely the “scale of likelihood of spontaneous occurrence” (Section 3).

2. Replacing ordering with distances

Scales in linguistics are normally conceived as a linear ordering of categories or functions. However, the restriction to a linear structure is neither necessary nor advantageous. Already the perennial issue whether first person should outrank second on the animacy scale (or vice versa) illustrates that a linear ordering is simply not powerful enough to model linguistic diversity. Probably the only reason for the existence of this focus on linear orders is that such scales are easier to handle and easier to visualize. Also, many scales currently being discussed only consist of two entities, and then the whole issue of linearity does not arise. However, to generalize the notion of a scale, it seems more fruitful to abandon the principle of a linear scale and open up the possibility for more complex topologies.

One proposal for more complex structures is implicit in the spider-web-like graphs used to display semantic maps (Croft, 2003, 133-139; Haspelmath, 2003). I have argued elsewhere that such semantic maps can be generalized as dissimilarity matrices (Cysouw, Ms.; Cysouw, 2007). In a dissimilarity matrix, all pairs of entities in the scale are considered

separately, and evaluated individually on their similarity. A linear scale is a special case of such a general structure. For example, consider three entities A, B, and C on a linear scale (in this order). This translates to similarities by stating that the distance from A to B is the same as the distance from B to C, and both are exactly half the distance from A to C. When the distance from A to C is not exactly matching the summed-up distance from A to B to C, the distances do not fit on a linear scale anymore. The principle of the generalization proposed here is that linearity might still exist, but it is not assumed *a priori*. Initially, the pairwise distances are established individually, and only afterwards it might turn out that they reduce to a nice linear arrangement. However, most of the times they actually will turn out not to be that easily aligned.

Often—though not in all instances—scales are considered to be inherently directed, i.e. they have a “high end” and a “low end”.² This direction is solely the result of the formal encoding of the scale (cf. Croft, 2003, 140-142). I will argue that any direction of a scale is an independent insight from any discovery of underlying topology of the scale. How to proceed with the separation of these two issues is the central proposal of this paper.

3. Scale of likelihood of spontaneous occurrence

As an example of the approach proposed here I will reanalyze 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 (i.e. an expression without an instigator of the action, like *the water boiled*), and a causative expression (i.e. an expression with an instigator, like *the man boiled the water*). He proposes the scale of “likelihood of spontaneous occurrence” to explain why—across languages—some predicates tend to be causativized (i.e. the causative is morphologically derived from the inchoative, e.g. German inchoative *enden* vs. causative *beenden*), while others tend to be

²The combination of a linear scale with an inherent direction is mathematically equivalent to a so-called “total order”. An easy way to remedy the problem of having parts of the scale that are not (clearly) ordered relative to each other, like first and second person, is to allow for some parallelism in the ordering (thus deviating from strict linearity). Such a model is mathematically speaking an example of a so-called “partial order”.

anticausativized (i.e. the inchoative is derived from the causative, e.g. English inchoative *be destroyed* vs. causative *destroy*). The idea of the scale of likelihood of spontaneous occurrence is that those predicates that are likely to occur spontaneously (i.e. without any human agent) will cross-linguistically tend to be causativized (i.e. the inchoative is the more basic expression), and *vice versa*.

Haspelmath investigated the inchoative/causative alternation of 31 meanings in 21 languages. For each of these meanings, the proportion of languages that use a causativizations strategy is shown in Table 1.³ The order of the meanings in the table illustrated the idea of the scale of likelihood of spontaneous occurrence, with the least spontaneous meanings being at the bottom (*split, close, break*) and the most spontaneous meaning being on top (*dry, freeze, boil, die/kill*).

Table 1: Cross-linguistic proportion of causativizations (adapted from Haspelmath 1993: 104)

split	0.04	develop	0.33	melt	0.68
close	0.06	roll	0.35	learn/teach	0.68
break	0.07	spread	0.35	sink	0.70
open	0.10	begin	0.38	go out/put out	0.71
gather	0.12	finish	0.38	wake up	0.75
change	0.12	fill	0.38	dry	0.77
connect	0.14	be destroyed/ destroy	0.39	freeze	0.86
rock	0.25	burn	0.42	boil	0.96
improve	0.26	dissolve	0.42	die/kill	1.00
rise/raise	0.27	turn	0.48		
get lost/ lose	0.28	stop	0.62		

³The proportion of causatives reported in Table 1 is calculated by dividing the number of languages that causativize the predicate (C) by the number of the languages that either causativize or anticausativize it (C+A), ignoring those languages that use different strategies (Haspelmath further distinguishes suppletive, labile and equipollent alternations, which will not be used here). This method of calculation is different from the proportions reported on by Haspelmath (who lists the fraction A/C), and is noteworthy in the case of *die/kill* as most languages use a suppletive strategy for this meaning, making the proportion reported here somewhat superficial (A=0 and C=3, so C/C+A=1).

4. Scales of functions

In most current research, scales of categories/functions are either available as hypotheses from earlier research, or established *post-hoc* as the most compelling way to interpret an observed cross-linguistic scale of form (and as such, they can be used as hypotheses for future research). However, scales of function can actually be established independently by using the semantic map approach. However, such “scales” will normally not be nicely one-dimensional, needing more effort for their interpretation.

The basic intuition behind the semantic map approach is that cross-linguistic variation in the expression of the functions can be used as a proxy to the relation between the functions themselves. The central assumption made in this approach is that when the expressions of two functions are similar in language after language, then the two functions 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 (see Cysouw, Ms. for a detailed exposition of this approach).

So, the similarity between expressions is the basic measure to establish a semantic map. However, there are two crucially different kinds of similarity between expressions, only one of which will be used to establish the semantic map. For a semantic map, it is important whether two expressions are similar on purely language-specific grounds, i.e. they behave alike according to the grammar of the language. For example, the English verbs *walk* and *enter* behave alike as to the formation of their past forms (*walked*, *entered*). Likewise, the verbs *buy* and *fight* behave alike in choosing the same kind of past formation (*bought*, *fought*). This kind of similarity between expressions is purely language-particular and thereby crucially different from cross-linguistic coding strategies. Cross-linguistically, one might say that *walk* and *enter* both use a suffixal concatenative morphology to mark past, but *buy* and *fight* use ablaut-like non-linear morphology. Such characteristics lead to scales of form to be discussed in the next section.

Coming back to the inchoative/causative alternation, the English expressions *open* and *close* have some similarity because they use the same inchoative/causative alternation (both verbs do not change their morphology in this alternation, i.e. they use a labile strategy in the terminology of Haspelmath 1993). Likewise, the German expressions *öffnen* and *schließen* are similar because they use the same inchoative/causative alternation,

though they use a different strategy from the one found in English (in German both verbs are anticausativizing, as the inchoative form is derived from the causative form by using reflexive morphology: *sich öffnen*, *sich schließen*).

By combining such language-particular similarities from many languages, the similarity between functions/meanings can be approximated. So, in the above example, both English and German use the same (language-particular) construction for “to open” and “to close”, so both languages argue for some similarity between these meanings. However, this is not necessary the case in all the world’s languages. For example, in Hindi the inchoative/causative alternation for the verb “to open” is coded by non-linear morphology (ablaut) *khulnaa/kholnaa*, but the alternation for the verb “to close” it is coded by using the copula-like verbs *hona* “to be” and *karna* “to do”, viz. *band honaa/band karna*, lit. “be open/do open”. Now, it is possible to approximate the similarity of the meanings “to open” and “to close” by (roughly speaking) taking the average of many such language-particular similarities.

When this procedure is followed for all possible pairs of meanings, this will result in a long list of similarity measures two meanings. For example, in the case of the meaning investigated by Haspelmath, there are $31 \cdot 30 / 2 = 465$ such pairs of meanings. Such a long list of numbers (a “dissimilarity matrix”) is a generalized scale of meaning. The network in Figure 1 is an attempt to display the structure of the resulting “scale” of meanings. The figure shows a so-called “splits graph” (Bandelt and Dress, 1992; Dress and Huson, 2004).⁴ Roughly speaking, similar functions will be placed close to each other in the network-like graph. At the upper right side of the figure some of the more spontaneous meanings can be found (e.g. *freeze*, *dry*, *burn*, *boil*, cf. Table 1), and at the opposite site, in the left and lower left, the meanings that typically need an agent are located (e.g. *break*, *split*, *open*, *change*). So, it is already possible to discern something like the scale of likelihood of spontaneous occurrence. However, this graph also very clearly shows that the spontaneity scale is not the only things that matters. As one might expect, meaning/function is a highly complex and multidimensional matter—and a multitude of other aspects of meaning are relevant for the similarity of meaning between the meanings investigated.

⁴The particular splits graph shown in Figure 1 is a NeighborNet made by the program SplitsTree (Huson and Bryant, 2006). See Bryant et al. (2005) for an introduction to this approach with some examples from linguistics.

Another way to depict the structure in the list of pairwise similarities is to use multidimensional scaling.⁵ Shown in Figure 2 are the first two dimensions of a multidimensional scaling for the same data leading to the network in Figure 1. Because only the first two dimensions are shown, this display might look easier to interpret, but that is only because much of the complexity of the data is ignored to fit the display into two dimensions. The spontaneity scale can be seen ranging from the upper left to the lower down side in Figure 2 (cf. Table 1). The meanings in the upper left of the figure are highly spontaneous (*boil, freeze, dry*), while the meaning at the lower right typically need an agent (*open, split, close, break, change*).

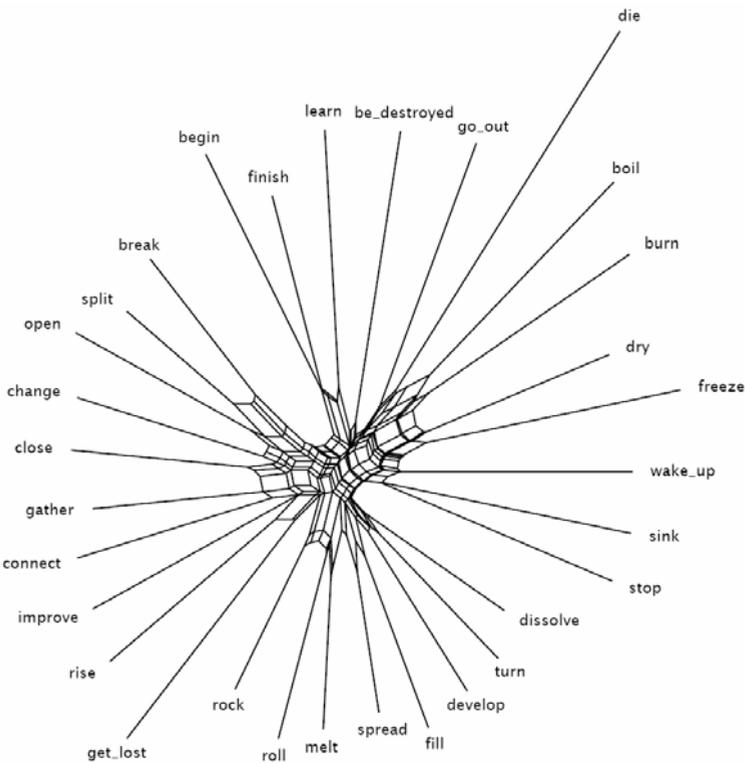


Fig. 1: A “scale” of functions, in the form of a NeighborNet of the 31 meanings.

⁵For all multidimensional scaling in this paper I used the function *cmdscale* from the statistical environment *R* (R Development Core Team, 2007).

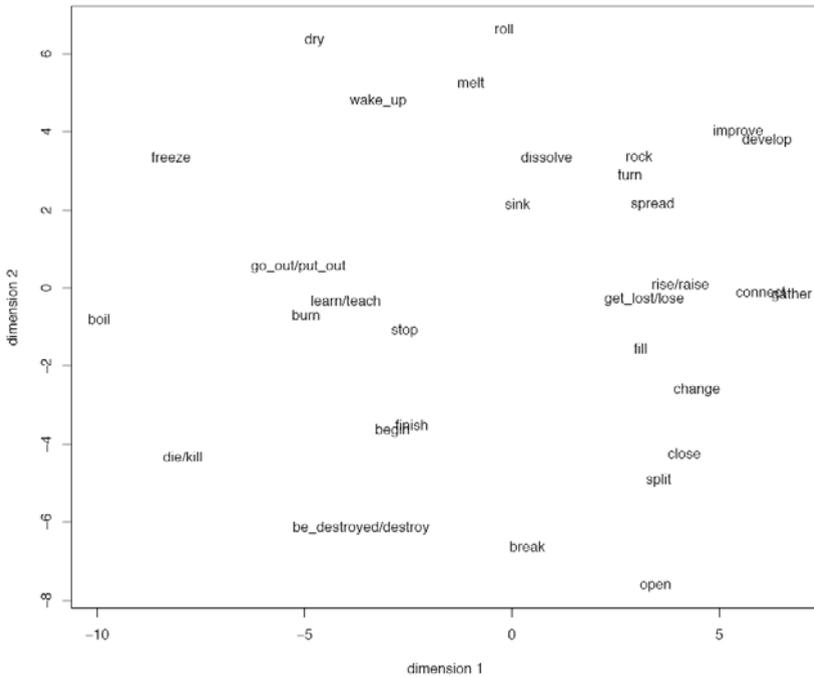


Fig. 2: Multidimensional scaling of the 31 meanings.

5. Scales of form

The constructions that languages use to mark the inchoative-causative alternation are not directly comparable across languages. Take, for example, the German construction using *sich* to mark the inchoative (e.g. *sich öffnen*, *sich schließen*). This construction is very reminiscent of the Hebrew construction using *hit-* to mark the inchoative (e.g. *hitʕorer*, *hitʕasef*). Both constructions are of course different in principle—after all, they come from different languages. However, there are various characteristics that make both constructions alike to some extent. For example, they both explicitly mark the inchoative in relation to the causative. Also, both constructions perform this marking by putting some extra material in front of the lexical verb (though there is of course a difference in morphological status). Further, both the German *sich* construction and the Hebrew *hit-*

construction are sometimes considered to be “reflexive” constructions. These three examples are characteristics of expressions that are cross-linguistically applicable, and in this sense crucially different from the characteristics that are used to establish language-specific construction classes.

Such cross-linguistically applicable characteristics of expressions are called “strategies” in the typological literature. This tradition to use the term “strategy” probably originated with Keenan & Comrie’s (1977, 64) classic paper on relativization strategies. There are different kinds of strategies, and these different kinds have a rather different status for the comparison of languages, but that topic will not further developed here. I will here only use so-called “coding” properties that relate to the form in which the language-particular expressions are codified.⁶ In this realm, one can think of characteristics like length of forms, kind of morphological process, or order of elements. The similarity of constructions with respect to such a coding property is here called a scale of form.

In Haspelmath’s original 1993 paper on the inchoative/causative alternation, he distinguishes five different coding strategies that languages use to mark the alternation: causative, anticausative, equipollent, labile, and suppletive. Causative constructions are inchoative/causative pairs in which the causative is morphologically overtly derived from the inchoative. Anticausative constructions are the opposite: the inchoative is overtly derived from the causative. Labile constructions are alternations that do not show any overt marking on both inchoative and causative, in contrast to equipollent constructions that have some marking on both. Finally, suppletive constructions are inchoative/causative alternations where there is no (obvious) morphological relation between the two forms. The central opposition in this scale of form is the causative vs. anticausative opposition, the analysis of which led Haspelmath to the spontaneity scale (cf. Table 1).

Many such scales of form can rather easily be approximated by automatically generated measures. Such measures will never be perfect from a linguist’s perspective, but they will get the job done much more quicker. For example, simply consider counting the number of unicode

⁶Besides coding properties, Keenan (1976) also distinguishes behavioral properties of expressions in complex constructions as another kind of constructions. Note that the German/Hebrew example discussed above included yet another kind of cross-linguistic strategy. The impression that both the German *sich* construction and the Hebrew *hit*-construction are “reflexive” constructions can be formalized by including reference to a “prototypical” element in the realm of meaning. For example, constructions from different languages are both reflexive-like when they both at least code for the meaning “rise”.

characters used in the written version of the inchoative and causative forms. A plot for the average word-length of the inchoative vs. the causative is shown in Figure 3. Obviously, these two counts are strongly correlated because in most cases there is regular morphology deriving one from the other, and the counts of characters include the length of the stem. However, there appears to be an interesting cross-linguistic cline in the total length of the meanings. The expressions of “die/kill” tend to be short, while the expression of “develop” and “improve” seem to be long across languages. This cline might be correlated to frequency of use, in that more frequently occurring meanings have shorter expressions. However, quickly checking some online corpora, the lower left to upper right cline in Figure 3 does not seem to correlate well with pure token frequency.

More relevant to the current topic is the upper left to lower right cline in Figure 3, which represents the different in length between the causative and the inchoative form. Meanings in which, across languages, the causative is longer than the inchoative should correspond to those meaning that have a preference for causativization, i.e. they should be high on the spontaneity scale shown in Table 1. And indeed, as shown in Figure 4 the spontaneity scale strongly correlates with the average causative-minus-inchoative character count ($r=.89$).⁷

⁷Actually, the only meaning clearly being off on this correlation is “die/kill”, which is probably an effect of the imprecise estimate on the spontaneity scale (cf. Footnote 3), and not so much an error of the approximation of counting characters. It does seem to make more sense to place “die/kill” somewhere on the higher middle of the spontaneity scale (as suggested by the counts of characters) than to place it completely on top (as suggested by the spontaneity scale in Table 1). After all, dying is indeed commonly a spontaneous activity, though it is not that uncommon to be induced by an agent.

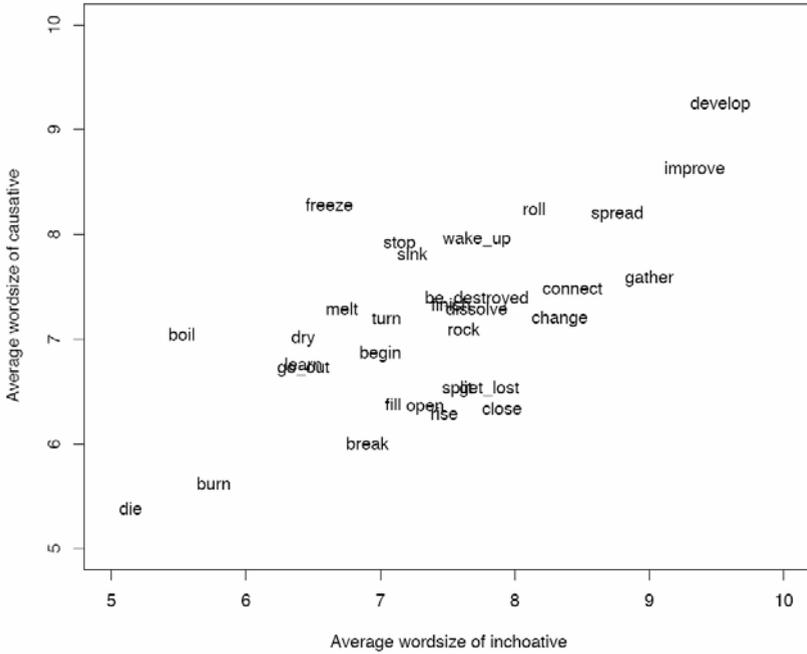


Fig. 3: Average number of characters used to mark inchoative (x-axis) and causative (y-axis).

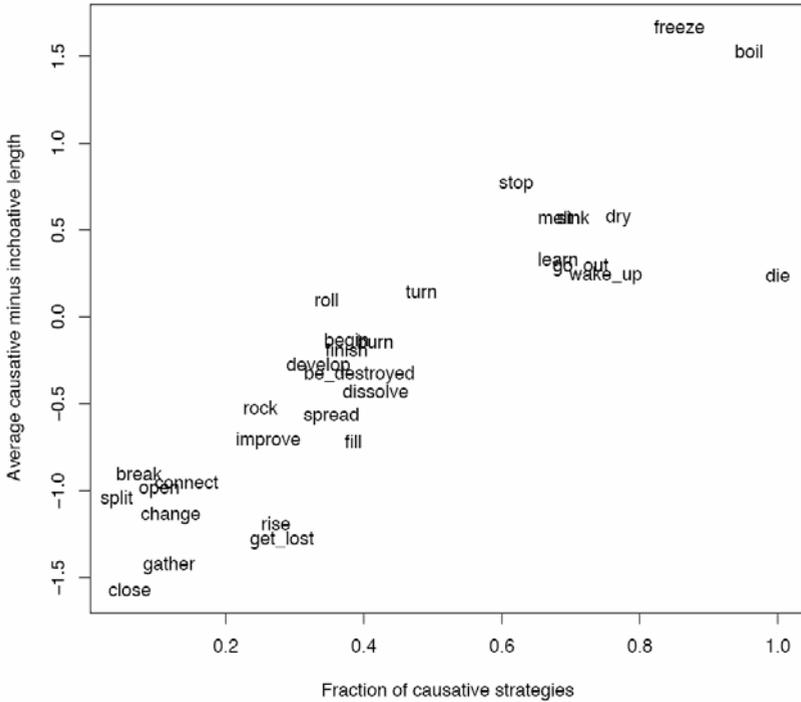


Fig. 4: Spontaneity scale approximated by average length difference of inchoate and causative form.

6. Matching form and function

Given a scale of function and a scale of form, the basic idea now is to investigate how these two scales match. In general, it is not immediately obvious how this should be done, but for specific cases there are many nice techniques to visualize such correlations and investigate their statistical significance. The examples discussed below illustrate some of the possibilities.

In Figure 5, the multidimensional scaling from Figure 2 is used as the basis to display the structure of the semantic map. As an overlay over this display the length difference of the forms is shown. This overlay is like a geographic map using contour lines (technically called “isohypses”) to indicate elevation. The level of elevation is defined by the total difference of characters throughout all 21 languages between the inchoative and

the scale of functions/meanings can be explained by the length of the forms. This problem is somewhat alike to a multivariate analysis of variance, weren't it for the fact that the variable to be explained (the scale of functions/meanings) is of a rather unusual kind, namely a dissimilarity matrix. Recent work in bioinformatics (Zapala and Schork, 2006) fortunately presents a solution for this particular problem.⁹ As shown in Table 2, the length of the causative and the length of the inchoative explain about 36% of the distances in the scale of function/meaning. Or, more to the point, the difference in length between inchoative and causative is explaining about 21% of the variation, and the sum 15% (as shown in Table 3).

Table 2: ANOVA of distance matrix by length of forms.

	Sums of Sqs	Mean Sqs	F Model	R2
length of causative	0.022	0.022	5.449	0.116 ***
length of inchoative	0.048	0.048	11.759	0.249 ***
combined effect	0.012	0.012	2.962	0.063
Residuals	0.111	0.004		0.572

Table 3: ANOVA of distance matrix by length difference and length sum.

	Sums of Sqs	Mean Sqs	F Model	R2
length difference	0.041	0.041	9.594	0.211 ***
length sum	0.030	0.030	6.922	0.152 ***
combined effect	0.008	0.008	1.758	0.039
Residuals	0.116	0.004		0.596

A different approach to the correlation between the scale of function and the scale of form is by using matrix correlation. Basically, the idea is to also consider the scale of form as a dissimilarity matrix and then correlate the form matrix with the function matrix. To reformulate the measurements of form (in the current case, these measurements are the average length of the inchoative and the causative expressions) into a dissimilarity matrix, all pairs of measurements have to be compared individually. As a dissimilarity one can, for example, simply take the Euclidean distance between the measurements for each pair. This dissimilarity in effect represents the linear distance between the words as shown in the configuration of Figure 3. The

⁹The multivariate ANOVAs shown in Table 2 and Table 3 were calculated by using the function *adonis* in the *R* package *vegan* (Oksanen et al., 2007).

length of a direct line between two words in that figure is the same as the Euclidean distance. Mathematically defined, this amounts to taking the dissimilarity as defined in:

$$d(A, B) = \sqrt{(Inch_A - Inch_B)^2 + (Caus_A - Caus_B)^2}$$

For example, “boil” has an average length of 5.52 characters for the inchoative and 7.05 for the causative. Likewise, “freeze” has an average length of 6.62 for the inchoative and 8.29 for the causative. Taking the Euclidean distance between the point (5.52, 7.05) and (6.62, 8.29) results in a dissimilarity between “boil” and “freeze” of 1.66. Doing these calculations for all pairs results in a dissimilarity matrix of form. Figure 6 shows the correlation between this scale of form and the scale of function. Each point in this figure represents one pair of meanings, plotting the dissimilarity of function against the dissimilarity of form. The figure already shows a rather nice correlation, which can also be shown to be statistically significant ($r=.40$, Mantel test $p<.0001$).¹⁰

¹⁰The Mantel test (Mantel, 1967) was performed using the function *mantel.test* in the R package *APE* (Paradis et al., 2004).

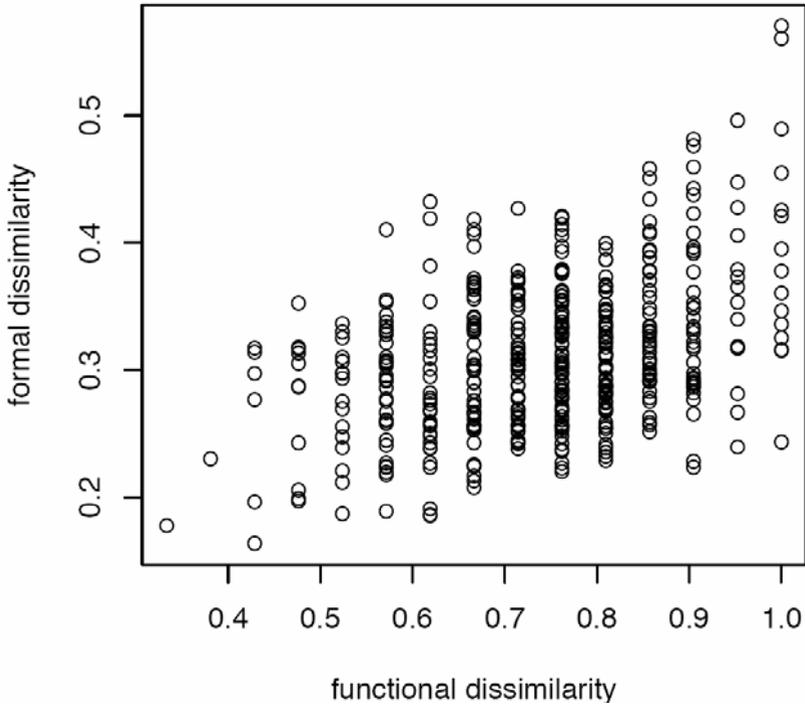


Fig. 6: Correlating form and function dissimilarities.

7. Conclusions

The following points summarize the proposals put forward in this paper:

- A scale consists of three parts: a scale of functions, a scale of form, and a match between the two.
- Both the scales of functions and the scales of form are not necessarily linear. They can be internally structured in complex ways, though even such complex structures are interesting and represent strong restrictions on possible linguistic variation.
- The most general description of the internal structure of these scales are dissimilarity matrices, which might boil down—under special circumstances—to be a description of a linear structure.
- The match between form and function is a kind of matrix correlation, though other methods might also be used. However, this is an area where

much work has to be done to elucidate which approaches are most suitable for linguistic typology.

References

- Bandelt, HJ, and AW Dress. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* 1, no. 3: 242-252.
- Bryant, David, Flavia Filimon, and Russell D. Gray. (2005). Untangling our past: Languages, trees, splits and networks. In *The Evolution of Cultural Diversity: A Phylogenetic Approach*, eds. Ruth Mace, Clare J. Holden, and Stephan Shennan, 67-84. London: UCL.
- Croft, William. (2003). *Typology and Universals*. Cambridge Textbooks in Linguistics). Cambridge University Press.
- Cysouw, Michael. (2007). Building semantic maps: the case of person marking. In *New Challenges in Typology*, eds. Bernhard Wälchli, and Matti Miestamo. Berlin: Mouton de Gruyter.
- Cysouw, Michael. Ms. Semantic maps as metrics on meaning.
- Dress, Andreas W. M., and Daniel H. Huson. (2004). Constructing Splits Graphs. *IEEE Transactions on Computational Biology And Bioinformatics* 1, no. 3: 109 - 115.
- Haspelmath, Martin. (1993). More on the typology of inchoative/causative verb alternations. In *Causatives and Transitivity*, eds. Bernard Comrie, and Maria Polinsky, 87-120. Amsterdam: Benjamins.
- Haspelmath, Martin. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, ed. Michael Tomasello, 2, 211-242. Mahwah, NJ: Erlbaum.
- Huson, Daniel H., and David Bryant. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23, no. 2: 254-267.
- Keenan, Edward L. (1976). Towards a universal definition of 'subject'. In *Subject and Topic*, ed. Charles N. Li, 303-333. New York, NY: Academic Press.
- Keenan, Edward L., and Bernard Comrie. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8, no. 1: 63-99.
- Mantel, Nathan. (1967). The detection of disease clustering and a

- generalized regression approach. *Cancer Research* 27, no. 2: 209-220.
- Oksanen, Jari, Roel Kindt, Pierre Legendre, Bob O'Hara, and M. Henry H. Stevens. (2007). vegan: Community Ecology Package. *R package*.
- Paradis, E., J. Claude, and K. Strimmer. (2004). APE: analyses of phylogenetics and evolution in {R} language. *Bioinformatics* 20, 289-290.
- R Development Core Team. (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Ribeiro Jr, Paulo Justiniano, and Peter J. Diggle. (2001). geoR: A package for geostatistical analysis. *R News* 1, no. 2: 15-18.
- Zapala, MA, and NJ Schork. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences of the United States of America* 103, no. 51: 19430-19435.