William Croft*

On two mathematical representations for "semantic maps"

https://doi.org/10.1515/zfs-2021-2040

Abstract: We describe two mathematical representations for what have come to be called "semantic maps", that is, representations of typological universals of linguistic co-expression with the aim of inferring similarity relations between concepts from those universals. The two mathematical representations are a graph structure and Euclidean space, the latter as inferred through multidimensional scaling. Graph structure representations come in two types. In both types, meanings are represented as vertices (nodes) and relations between meanings as edges (links). One representation is a pairwise co-expression graph, which represents all pairwise co-expression relations as edges in the graph; an example is CLICS. The other is a minimally connected co-expression graph – the "classic semantic map". This represents only the edges necessary to maintain connectivity, that is, the principle that all the meanings expressed by a single form make up a connected subgraph of the whole graph. The Euclidean space represents meanings as points, and relations as Euclidean distance between points, in a specified number of spatial dimensions. We focus on the proper interpretation of both types of representations, algorithms for constructing the representations, measuring the goodness of fit of the representations to the data, and balancing goodness of fit with informativeness of the representation.

Keywords: semantic map, co-expression, graph structure, Euclidean space, multidimensional scaling

1 Introduction

In this article I focus on two mathematical representations that have been used for what have come to be called "semantic maps" in typology. The two mathematical representations are graph structure and Euclidean space. In this section, I will provide a more precise characterization of "semantic maps" in typology, in order to place the mathematical representations and their use in context.

***Corresponding author: William Croft,** Department of Linguistics, University of New Mexico, Albuquerque, NM, USA, e-mail: wcroft@unm.edu

"Semantic maps" are a means for representing typological universals of linguistic co-expression across languages, with the aim of inferring similarity relations between concepts from those universals (for surveys and history, see Haspelmath 2003; Georgakopoulos and Polis 2018; 2021). As such, "semantic maps" function as one part of the process of doing typology: typological classification, typological generalization, and (functional-)typological explanation (Croft 2003: 1–2). "Semantic maps" are used for typological generalization. However, some clarification about the first and third steps in doing typology is necessary in order to understand how graph structure or Euclidean spatial representations can be used for typological research.

The type of data that is collected and classified for the purpose of constructing "semantic maps" is illustrated in (1) (Haspelmath 1997: 46, 42) and (2):

(1) a. *Masha met with **some**one at the university*. [speaker knows who it was]
b. *Masha met with **some**one at the university*. [speaker doesn't know who it was]
c. *Visit me **some**time*. [unspecified future time]

(2) a. *She **knows** Marvin*. [acquainted with]
b. *She **knows** that Marvin is Greek*. [factual knowledge]

In both (1) and (2), a single linguistic form is used to express two or more different meanings: the indefinite marker *some-* in (1) and the lexical verb *know* in (2). This phenomenon is most generally called *co-expression* (see for example Hartmann et al. 2014; Croft to appear, § 1.4). Both "grammatical" meanings (1-a)–(1-c) and "lexical" meanings (2-a)–(2-b) can be co-expressed, and generalizations for both are represented by "semantic maps" – either graph structures or Euclidean spaces.

In some analyses, such as Haspelmath's, the meanings expressed in the examples are intended to be representative of a more general semantic category: in the case of (1-a)–(1-c), specific known (referent), specific unknown and irrealis nonspecific respectively. In other studies, the semantics of a specific example are not generalized. For example, the co–expression of spatial relations elicited by the pictures in the Bowerman-Pederson picture set (see Levinson et al. 2003 and Section 3 below) apply only to the specific exemplar of a spatial relation in the picture (dog in a doghouse, necklace around a woman's neck, etc.). Some scholars have associated graph structure representations with general semantic categories and Euclidean spatial representations with exemplar data (e. g., Wälchli and Cysouw 2012: 679; Georgakopoulos and Polis 2018: 17). But classification of the data as exemplars or as more general semantic categories is independent of

the mathematical representation, if not the underlying semantic theory (see Section 7).

On the form side, co-expression is generally assumed to be phonologically identical, such as *some-* in (1) and *know* in (2), disregarding allophonic or allomorphic variation. Morphologically related but distinct forms such as English *hand* and *handle* are not considered to be co-expression in this strict sense. It is of course possible to use a looser definition of formal co-expression.

Most examples of data collection begin with a set of meanings or concepts, and examine all forms expressing those meaning for co-expression. This is an onomasiological approach (Georgakopoulos and Polis 2018: 5). Some studies then proceed to add other meanings expressed by the forms found with the original set of meanings (e. g., Youn et al. 2016; Georgakopoulos et al. 2021; Georgakopoulos and Polis 2021). However, this supplemental semasiological approach (Georgakopoulos and Polis 2018: 5) will miss co-expression relations among the added meanings that are not shared with the original meanings, and hence the structure of the network with the added meanings will be incomplete.

The classification and typological comparison of co-expression data described here is generally done with the aim of inferring similarity relations between the meanings or concepts that are co-expressed (or not co-expressed); I focus on that goal here. However, co-expression of meanings within or across languages may arise for reasons other than conceptual similarity.

Within a single language, two meanings may come to have the same form due to convergence of two phonological forms, as with English *to, too*, and *two*, whose independent historical sources are reflected in contemporary spelling. But for the vast majority of the world's languages, there is no direct documentation of a language's history. So another method must be used to distinguish accidental convergence from semantic similarity. This is where typology comes in. If two meanings are regularly co-expressed in a broad sample of languages, then the likelihood of co-expression being due to chance becomes vanishingly small: 'recurrent similarity of form must reflect similarity in meaning' (Haiman 1985: 26).

Comparing co-expression of meaning across languages brings up other potential explanations for co-expression across languages than similarity in meaning. The co-expression of meanings may be due to inheritance of the co-expression pattern from a common ancestor language. The co-expression of meanings may also be due to language contact, where speakers of one language copy the co-expression pattern from speakers of a neighboring language (see, e. g. van der Auwera 2013: 161).

These two possible historical explanations can be dealt with in the usual way in typology: by constructing a genetically and geographically stratified sample of languages (Youn et al. 2016 test their sample for these and other potential biases).

As with homonymy, using a stratified cross-linguistic sample does not a priori rule out accidental convergence, common ancestry or contact as reasons for particular co-expressions in the data. It reduces the likelihood of these explanations for the great majority of the data in the dataset (see Section 5), and allows one to focus on the patterns representing typological universals.

A different sort of historical explanation is relevant to the typological universals represented by "semantic maps" and their explanation in terms of conceptual similarity. It is hypothesized that co-expression arises when a form expressing a particular meaning is extended, or recruited, to express a related meaning. It is also sometimes hypothesized that this semantic extension or recruitment is unidirectional. For example, it is often assumed that forms for nonspatial meanings of adpositions are recruited from the forms used for spatial meanings (or, forms for spatial meanings are extended to non-spatial meanings). The semantic shifts in grammaticalization are also often assumed to be unidirectional (from "lexical" to "grammatical" meanings). We will return to this diachronic aspect of "semantic maps" in Section 2.

Once we have ruled out accidental convergence, common ancestry and contact in the usual typological way, there remains the question of what counts as semantic or conceptual "similarity". A better term to describe the range of semantic explanations would probably be semantic relatedness. Semantic extension may be attributed not just to similarity in the usual sense – similarity in certain semantic properties found in a single semantic domain – but also to metaphor and metonymy. Some have assumed that "conceptual similarity" should be construed narrowly, to exclude other types of semantic shifts such as metaphor and metonymy (Cristofaro 2010). I consider this to be too narrow a view (Croft 2010b). The aim is to uncover typological universals of semantic relatedness of concepts. A further level of explanation may be able to attribute the universals of semantic relatedness to different types of semantic shift.

The rest of this article will focus on using graph structure or Euclidean space representations for typological universals of linguistic co-expression. The use of these two mathematical representations are not as different as they are made out to be. Nevertheless, there is an interesting difference between graph structure and Euclidean space representations for one's theory of semantics that may influence one's choice of representation (see Section 7).

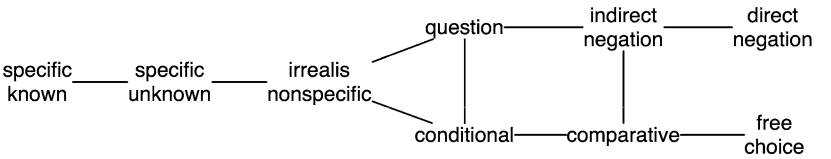

Figure 1: Semantic map of indefinite pronoun functions.

2 Graph structure representation: Basics

The graph structure representation is the more widely used representation in the typology of co-expression. An oft-cited example is that of indefinite pronoun meanings (Haspelmath 1997: 64), given in Figure 1.

Figure 1 is a *graph structure*: it is made up of *vertices*, also called *nodes*, given by labels in the figure ('specific known', 'specific unknown', etc.); and *edges*, the links between nodes in the figure. In the use of graph structure for representing typological universals of co-expression, individual meanings are represented as vertices, and the edges represent certain co-expression relations between two meanings.

Figure 1 represents a generalization over language-specific co-expression phenomena. Any particular indefinite pronoun in a particular language must map onto a connected subgraph of the graph. This property is what I called the Semantic Connectivity Hypothesis (Croft 2001: 96; 2003: 134). The Semantic Connectivity Hypothesis is one typological universal of co-expression that is represented by the graph in Figure 1. For example, English *some-* maps onto specific known, specific unknown, irrealis nonspecific, question and conditional (Haspelmath 1997: 65; see van der Auwera and Van Alsenoy 2011 for discussion of the semantics of the irrealis, question and conditional nodes). These five meanings form a connected subgraph. The subgraph is usually represented visually in publications by drawing a shape that includes the nodes in the subgraph and labeling this shape by the language-specific form.

At this point, there is some terminological ambiguity in the literature. The term 'semantic map' is used for both the entire graph structure in Figure 1, which is intended to represent certain typological universals, and a connected subgraph that a language-specific form such as English *some-* maps onto. I proposed describing the typological universal graph in Figure 1 as a 'conceptual space', and the mapping of language-specific forms as a 'semantic map' (Croft 2001: 92–95; 2003: 133–137). There is no consensus about terms for the typological universal graph vs. a language-specific form's subgraph, as far as I can tell. I will henceforth

use the terms *semantic map* and *conceptual space* as I have distinguished them in previous publications (and hence drop the scare quotes around "semantic map").

This is actually not a complete description of Figure 1, because not all co-expression relations are directly represented. Consider the English examples in (1-a)–(1-c). The morpheme *some-* is used for specific known and specific unknown meanings and specific unknown and irrealis nonspecific meanings, as indicated in Figure 1. As a result, there is also a co-expression relation between the specific known and irrealis nonspecific meanings – both are also expressed by *some-*. But there is no edge connecting the specific known node and the irrealis nonspecific node in Figure 1.

This is not an accident, but before explaining why, it should be noted that some typologists have used graph structures for lexical co-expression that do have edges for every pairwise co-expression relation between lexical concepts. These *pairwise co-expression graphs* are found, for example, in the Database of Cross-Linguistic Colexifications (CLICS; List et al. 2013; Rzymski et al. 2020; Jackson et al. 2019), a very large database of lexical co-expressions.

Why does Figure 1 lack an edge between the specific known and irrealis nonspecific nodes, as well as other nodes? The absence of that edge encodes an implicational universal: *if a language form co-expresses specific known and irrealis nonspecific indefinite pronoun meanings, then it also co-expresses the specific unknown meaning with these two meanings*. In other words, Figure 1 represents that the specific unknown meaning is "in between" the specific known and irrealis nonspecific meanings (cf. van der Auwera 2013: 155).

This is a more constrained representation of semantic relations (similarity) as inferred from co-expression data than what is found in a pairwise co-expression graph. Figure 1 is intended to represent the minimum number of edges required to maintain semantic connectivity (Georgakopoulos and Polis 2018: 6). I will call this a *minimally connected co-expression graph*. One important feature of a minimally connected co-expression graph is that it encodes information about the semantic structure of language-specific forms that map onto more than two nodes of the graph. For example, a minimally connected co-expression graph encodes that a three-way co-expression involving the specific known indefinite meaning must also involve specific unknown and irrealis nonspecific meanings, and so on. This information is not represented in a pairwise co-expression graph.

A final and important point about a co-expression graph is that the geometric arrangement of the nodes and edges on the page or screen is entirely irrelevant to the semantic relations represented in the graph (Haspelmath 2003: 233; Croft and Poole 2008: 6). The depiction of the graph on the page is purely a matter of convenience (avoiding crossing lines for the edges, for example). No inferences about semantic relations can be made from the positions on the nodes on the page. All

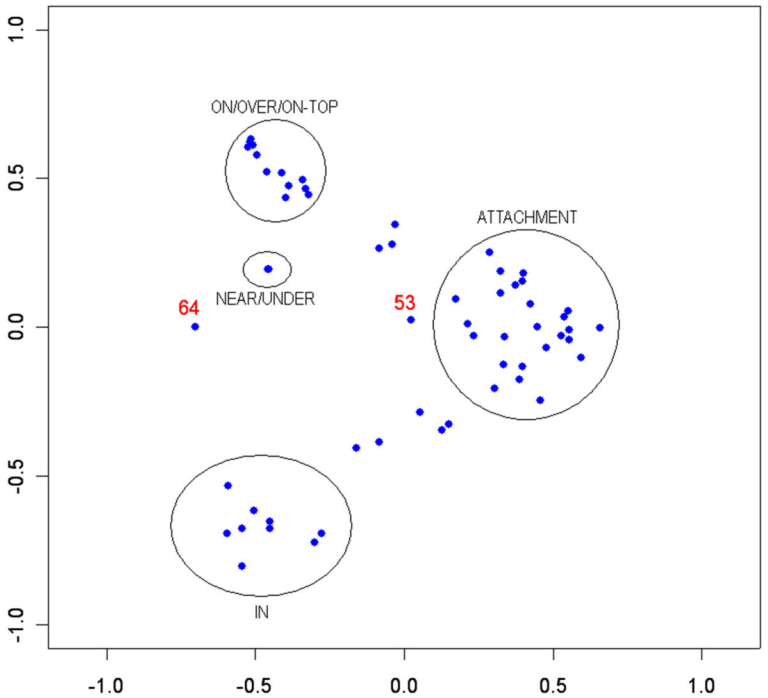

Figure 2: Two-dimensional multidimensional scaling model of adpositional spatial relations by unfolding.

that matters is whether two nodes are linked by an edge or not, not the positions of the nodes on the page. A graph structure representation is not a Euclidean space.

3 Euclidean space representation: Basics

An example of a Euclidean space representation is the two-dimensional spatial model of adpositional spatial relations in the Bowerman-Pederson picture set given in Figure 2 (Croft 2010a; data from Levinson et al. 2003, with thanks to Sérgio Meira).

Figure 2 is a two-dimensional Euclidean space, with points positioned in the two continuous dimensions. The points displayed in Figure 2 represent the meanings. The points are positioned in the two-dimensional Euclidean space such that *distance in any direction* represents degree of similarity/relatedness of the meanings, as determined by the typological co-expression data used to construct the

Euclidean space representation. A Euclidean space representation may have any number of dimensions, from 1 up to the number of co-expression relations in the dataset; see Sections 4–5 for the choice of the number of dimensions in a Euclidean space representation.

Figure 2 is a representation of the conceptual space. In a Euclidean space representation, language-specific categories – semantic maps, in the narrow sense given in Section 2 – are *bisections of the Euclidean space* (Croft and Poole 2008: 9–11). In a two-dimensional Euclidean space such as Figure 2, a bisection of the space is a straight line. In a three-dimensional Euclidean space, a bisection is a plane. In a one-dimensional Euclidean space, i. e. a linear representation, a bisection is a point that divides the linear conceptual space in half.

This representation of a semantic map in an MDS spatial model is sometimes misunderstood. A language-specific category in a Euclidean space representation is sometimes represented as a shape enclosing a set of points in the spatial model (e. g., Majid et al. 2011: 9 [Figure 7]; Hartmann et al. 2014: 472 [Figure 4]).

This practice may be carried over from the graph structure representation. Encircling the nodes that a language-specific form maps onto is fine in the graph structure representation. All that matters in the graph structure representation is that the language-specific category is a connected subgraph of the graph of the conceptual space. Since the graph structure is not a Euclidean structure, one can draw a line of any shape around the subgraph in order to show the reader that the subgraph is a language-specific category.

However, drawing a line of any shape around a set of points in a Euclidean spatial model constructed by multidimensional scaling is an incorrect representation of the semantic map of a language-specific category. Only a bisection of the space is a correct representation of the semantic map. Not recognizing this fact is possibly the reason for false statements such as that by Malchukov (2010: 177) that a Euclidean space representation is vacuous. (The shapes that encircle points in Figure 2 are intended to identify spatial meanings, not language-specific form categories; see also Levinson et al. 2003: 505–9.)

The fact that semantic maps are bisections of the space in a Euclidean space representation accounts for certain types of patterns that are found in Euclidean space representations of typological co-expression data. For example, a one-dimensional space can only represent a typological hierarchy (chain of implicational universals) where the language-specific categories always extend to one or the other end of the linear representation. The cutoff point bisects the linear hierarchy. This is the case with the animacy hierarchy mapped by occurrence of a distinct plural number form for nouns, which (almost) always includes the upper end of the hierarchy (Corbett 2000).

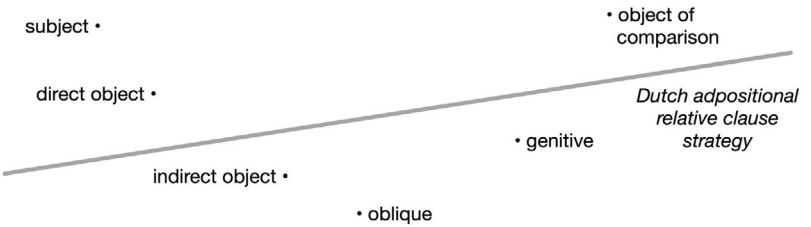

Figure 3: A semantic map of the Dutch adpositional relative clause strategy as a bisection of a Euclidean two-dimensional space. The two-dimensional space constitutes a representation of the relative clause accessibility hierarchy as a curved horseshoe shape, which allows the representation of a semantic map for a form that co-expresses functions in the middle of the hierarchy as a bisection of the space.

However, some hierarchies allow for categories that map into only a middle segment of the hierarchy. For example, there may be a relative clause accessibility strategy that applies only to the middle grammatical roles of the accessibility/grammatical relations hierarchy. The relative clause construction in Dutch with a relative pronoun indicating case (usually by an adposition) is used only for indirect object, oblique and genitive roles, according to Keenan and Comrie (1977: 76; 1979: 335). It is not used for either subject or direct object roles at the top of the hierarchy, or for the object of comparison role at the bottom of the hierarchy.

There cannot be a point that bisects a linear spatial model in such a way that the functions co-expressed by the form are on one side and the functions not expressed by the form are on the other side for the Dutch relative clause construction. However, a two-dimensional model can capture this with a curved line or 'horseshoe'. A straight line can bisect a curved representation so that the middle points representing functions co-expressed by the form (indirect object, oblique, genitive) are on one side, and the end points at either end (subject and direct object, and object of comparison), that is, the ends of the horseshoe, are on the other side; see Figure 3.

The horseshoe representation illustrated in Figure 3 is not an uncommon phenomenon in language universals. Examples of horseshoe-shaped patterns include the indefinite pronoun space corresponding to Figure 1 (see Croft and Poole 2008: 15 [Figure 4]), the object-property-event space for parts of speech (Rogers 2016: 74 [Figure 6]) and the 'on'-'in' spatial relation continuum in Figure 2 (Croft 2010a: 9 [Figure 4]; the ends of the horseshoe are the 'on' points at the top left and the 'in' points at the bottom left).

A rarer pattern is when the ends of a linear/hierarchical structure can "join up", leading to a circular representation. This is found in the co-expression

of lexical aspectual types in tense-aspect constructions (Croft 2012: 166 [Figure 4.4]).

Finally, a two-dimensional spatial model may represent a fully two-dimensional conceptual space with orthogonal (perpendicular) dimensions. For example, the spatial model for tense-aspect constructions generated from Dahl's (1985) questionnaire data has orthogonal dimensions for time reference (past-present-future) and aspect (imperfective-perfective; see Croft and Poole 2008: 26, Figure 8). The spatial model for encoding grammatical roles (case marking and indexation) has orthogonal dimensions for core vs. oblique participant roles and causally antecedent vs. subsequent roles (Hartmann et al. 2014; Croft to appear, § 6.1.2, Figure 6.1). The spatial model for constructions encoding the functionally related thetic, mirative and exclamative functions has orthogonal dimensions corresponding to entity-central vs. event-central functions (Sasse 1987) vs. stages in the psychological theory of surprise (García Macías 2016).

4 Algorithmic derivation of the typological universals

Finding the smallest number of edges that capture all the language-specific categories as connected subgraphs is not simple. Haspelmath (1997) constructed Figure 1 manually. Croft and Poole (2008: 6) observe that the graph in Figure 1 actually has an "unnecessary" edge, the edge between 'irrealis nonspecific' and 'conditional'. I observed this by manual re-inspection of the data. But for larger or more complex datasets, determining the smallest number of edges needed becomes impossible to do manually. In fact, this task is similar to the traveling salesman problem, which is known to be NP-hard (Croft and Poole 2008: 7).

Regier et al. (2013: 93–94) recognize this problem, and introduce an algorithm from the epidemiological literature that provides an efficient approximation for solving the problem of inferring the minimally connected co-expression graph with the minimum number of edges for a set of data.

The algorithm adds an edge to the graph containing the nodes of the concepts (or whatever entities are being characterized by similarity) in accordance to their utility, roughly, the extent to which the edge contributes to the goal of capturing the language-specific categories in the data as connected subgraphs of the eventual conceptual space graph (see Regier et al. 2013: 94–95 for details).

Regier et al. apply the algorithm to the same two datasets discussed above: indefinite pronouns from Haspelmath (1997), and spatial relations from Levinson et al. (2003). They find that the graph from Figure 1 minus the edge from 'irre-

alis nonspecific' to 'conditional' is indeed the simplest graph that captures all of Haspelmath's data as connected subgraphs.

Regier et al. then apply the algorithm to the much more complex data of spatial relations. The data involve 71 concepts – the Bowerman-Pederson picture set described in Section 1 – instead of just nine; and adposition categories from nine languages. The graph structure is displayed as Figure 5 in their article (Regier et al. 2013: 100).

Euclidean space representations are constructed by an algorithm for even the smallest datasets. MDS spatial representations for typological universals of coexpression can be generated by at least two different types of algorithms. In the more common type, dissimilarity (used for example by Levinson et al. 2003 and Wälchli and Cysouw 2012: 680–681), one looks at pairs of concepts and scores how frequently the pair of concepts are expressed by the same form (with a number between 0 and 1). Thus, raw distributional data – for each form, what concepts it expresses and what concepts it does not express – must be converted into a square matrix comparing each pair of concepts, and a calculation of how many forms in the data express both concepts.

The other type of algorithm, unfolding, takes the distributional data directly (Poole 2000; Croft and Poole 2008). That is, one creates a matrix of distributional data, with the concepts in the rows and the forms in the data in the columns. This is a rectangular matrix: the number of concepts being analyzed may be different from the number of forms in the data. Each cell in the matrix scores whether the concept is expressed by the form or not. Since similarity is represented as small distance, Poole's unfolding algorithm scores expression as 1 and nonexpression as 6; missing data is scored as 9.

The unfolding algorithm is superior to the dissimilarity algorithm if the data is lopsided, that is, there are many categories expressing either just a few concepts, or expressing most of the concepts (a very general category). Dissimilarity will treat the former as extremely dissimilar and the latter as extremely similar. This compresses the numerical range of values and magnifies random processes due to noise. Levinson et al.'s spatial adposition data is very lopsided in both ways. The spatial model produced by the unfolding algorithm has a much more coherent semantic interpretation than the spatial model produced by the dissimilarity algorithm (see Croft 2010a: 8–10 for more discussion).[1]

1 Poole, who developed the unfolding algorithm for the analysis of voting patterns in legislatures, has more recently developed a Bayesian algorithm for constructing Euclidean space representations of voting patterns (Bakker and Poole 2013). This algorithm could also be adapted for typological analysis.

5 Goodness of fit in graph structure and Euclidean space representations of typological universals of co-expression

It is often said that the only exceptionless typological universal is that all typological universals have exceptions. From Greenberg (1966), who qualified his implicational universals with phrases such as 'almost always' and 'with overwhelmingly greater than chance frequency', to the present, typologists have acknowledged that the data supporting universals is complex, messy and noisy. Some typologists try to explore in depth the "exceptions" to their proposed universals, in order to find a diachronic or other explanation for the anomalies (e. g., Stassen 1997; 2009). In some cases, such searches are unsuccessful, not least because of the incompleteness or vagueness of the data. Others have followed other empirical sciences by developing statistical tests to confirm or disconfirm typological universals induced from cross-linguistic data, such as Maslova's (2003) significance tests for implicational universals.

Users of minimally connected co-expression graphs have generally not adopted this approach. Every instance of co-expression in the cross-linguistic data is given an edge in the graph structure. The result leads to a graph structure that conforms to the Semantic Map Connectivity Hypothesis and the minimal connection criterion without exception. But the graph structure that results may be a rather weak model of semantic similarity/relatedness, despite imposition of the similarity requirement. It may also be that the rare or otherwise less significant co-expression patterns are actually not reflecting semantic relatedness. Instead, they may reflect accidental convergence, common ancestry, language contact, and possibly other language processes. This is a criticism of the graph structure model for semantic similarity by Cristofaro (2010) and others. After all, as noted in Section 1, the typological method does not eliminate these alternative explanations; it only reduces their likelihood. And likelihood is statistical.[2]

All quantitative models of data balance informativeness of the model – what broad patterns it reveals – and accuracy – how much of the data is predicted cor-

2 Van der Auwera avoids the problem of alternative explanations than semantic relatedness by using only data of attested semantic shift processes (van der Auwera and Plungian 1998; van der Auwera et al. 2009), which are represented as directed edges (links) between meanings in the graph structure. (He also categorizes edges by the type of semantic shift; van der Auwera 2013: 161.) However, this greatly limits the data that can be used. As a result, it most likely underestimates the possible paths of semantic shift (see the added paths in the 2009 publication), and does not provide a cross-linguistically accurate measure of frequency.

rectly by the model. This is the consequence of the signal vs. noise issue raised in Section 1. Linguistic behavior is extremely complex. Many different factors play a role in how speakers express concepts in linguistic form, and in how linguists collect and analyze data for their theories. Some of these factors may be stochastic, and may compete with each other. Random errors in data collection and analysis also occur. Thus, a model that has a perfect fit to the data may not actually be that informative (Levinson et al. 2003: 499, fn. 7; Croft and Poole 2008: 11–12).

In other words, goodness of fit must always be addressed in constructing either graph structure or Euclidean space representations of typological universals of co-expression. Goodness of fit allows us to get an idea of the tradeoff between how much of the cross-linguistic data our representational model captures, and how useful the model is for inferring conceptual similarity/relations, the phenomenon we are most interested in explaining with the co-expression data.

For example, MDS represents similarity data as Euclidean distance. Concepts may be similar to each other in different ways, or for different (semantic) reasons. These different ways can be captured by different dimensions in a spatial model. MDS does this by capturing *all of the variance* in the data in the number of dimensions used in a particular model.

If one limits the MDS model to one, two or three dimensions, or even more dimensions, one has constrained how many different co-expression patterns can be captured. In a low-dimensional representation, there will not be a perfect fit of the model to the data: that is, some points will fall on the "wrong side" of the cutting line separating concepts expressed by the form from concepts not expressed by the form. One could add another dimension to the spatial model in order to make the model fit the data better; but that makes the model "looser", adding another explanatory dimension. One is guaranteed a perfect fit by adding as many dimensions as there are linguistic categories. But such a model is completely uninformative, capturing no generalizations at all.

Thus, one must provide a measure of goodness of fit, and compare the goodness of fit of related models. This is rarely done in linguistic applications of multidimensional scaling, let alone graph structure representations. In MDS, the related models to be compared are models in one, two, three, or more dimensions. For example, Table 1 provides two goodness of fit statistics for MDS models by unfolding in one, two and three dimensions for the spatial adposition data from Levinson et al. 2003 (Croft 2010a: 9).

The percent correct classification is the percentage of points on the "right side" of all the cutting lines (semantic maps) in the data. (Poole's unfolding algorithm maximizes correct classification; Croft and Poole 2008: 18.) The aggregate proportional reduction of error is a measure of how much the model is an

Table 1: Fitness statistics for the spatial adposition co-expression data in Levinson et al. (2003) for MDS models by unfolding (see Figure 2 and Croft 2010a).

Number of dimensions	% Correct Classification	APRE*
1	94.1	.300
2	95.8	.501
3	97.1	.661

*aggregate proportional reduction of error (see text)

improvement in accuracy over a null model (in which the cutting lines either include all points or exclude all points; the APREs for this data are relatively low because the data is so lopsided). Adding dimensions to the spatial model improves correct classification and APRE, but adding the third dimension improves fit to the data less than adding the second dimension does, particularly for the APRE. For this reason, the two-dimensional model is chosen as providing the best balance between informativeness and accuracy. But it assumes that there is some inaccuracy, essentially, "noise", in its effort to capture a signal apparent in the two-dimensional spatial representation.

One can also create goodness of fit statistics for graph structure representations. Even a minimally connected co-expression graph may have a large number of edges. For complex datasets with many different concepts such as the Levinson et al. data, a minimally connected graph with a 100 % fit to the data is quite difficult to interpret. Regier et al.'s minimally connected graph for adposition co-expression (their Figure 5) is very densely connected, with many crossing lines in the two-dimensional representation necessary on a printed page.

One can add a goodness of fit statistic to a minimally connected co-expression graph based on the utility function. One can then prune the least informative edges, that is, the edges with the lowest utility score(s). The program used by Regier et al. outputs a list of edges and their statistics, including the utility score of the edge. The edges are added by their utility score ranking; so the last edges added have a utility score of only 1. Table 2 gives the number of edges added by utility score for the Levinson et al. spatial relation data (with thanks to Terry Regier for providing the utility score data).

It can be seen that there is a great leap in the number of edges added for the edges with the lowest utility score: 37 of the 115 edges have a utility score of only 1. One could prune one third of the edges without a great loss of utility. The resulting graph would be more easily interpretable – though, admittedly, I have not constructed this graph so I do not know how interpretable it is in comparison to the two-dimensional MDS spatial model of the same data. More generally, mak-

Table 2: Number of edges added to the graph for the conceptual space of adpositional spatial relations, by their utility score.

Utility score (10 = highest)	Number of edges of this utility score added to the graph
10	2
9	11
8	3
7	3
6	6
5	13
4	15
3	10
2	14
1	37

ing the graph structure model more informative by using fitness statistics to prune the least informative edges, comparable to reducing the dimensionality of an MDS spatial model, might make more complex graph structure representations more interpretable.

Georgakopoulos and Polis (2021) independently developed the idea of using fitness statistics to prune low-utility edges. They also weight the edges that are produced by the algorithm for the minimally connected graph – the edges that are minimally necessary to satisfy connectivity, minus the edges pruned due to low utility score – visually by thickness.[3]

Weighted edges, visualized by edge thickness, are also used in the pairwise co-expression graphs such as those found in Cysouw (2007: 232–234), Youn et al. (2016) and in CLICS (List et al. 2013; Jackson et al. 2019; Georgakopoulos et al. 2021). Edge weight/thickness is use to capture relative likelihood of co-expression in languages, which in turn is hypothesized to reflect degree of semantic similarity or relatedness. List et al. (2013: 349) also suggest balancing goodness of fit to informativeness of the graph structure representation by pruning the lowest weight edges in a pairwise co-expression graph.[4]

3 The utility score is not always the same as the number of forms that co-express the two meanings linked by the edge. The utility score is recalculated when the highest-utility edge(s) is/are added to the minimally-connected graph being constructed, and that score may change and hence diverge from the number of forms co-expressing the two functions. (Thanks to Terry Regier for clarifying this point.)

4 However, if lower-utility edges are pruned from a minimally-connected graph, then some nodes – that is, some meanings – will not be connected to the graph; or in a weighted-edge rep-

Pruning of edges would help to address the issue of idiosyncratic co-expression patterns not likely to be due to semantic similarity or relatedness. However, relative weighting of edges is indicative of likelihood of co-expression and semantic relatedness most clearly in a genetically and geographically stratified sample, such as the one used in Youn et al. (2016). CLICS is not based on such a sample, so relative weight of edges in a CLICS graph will reflect sampling bias as well as likelihood of co-expression. Finally, simple Euclidean distance in a Euclidean space representation captures more directly the likelihood of co-expression/degree of semantic relatedness of meanings or concepts.

6 Other mathematical models related to multidimensional scaling

Multidimensional scaling is one member of a family of models for multivariate analysis. Other models in this family that have been used in linguistics or linguistic typology are principal component analysis, factor analysis and correspondence analysis. These other models differ from multidimensional scaling in certain ways.

Multidimensional scaling is an unsupervised distance model. An *unsupervised* model is one where the categories or groupings (the clusters observed in Figure 2, for example) are not specified in advance. In this respect, MDS differs from some cluster analysis models where the number of clusters must be specified in advance.

A *distance* model is one in which Euclidean spatial distance represents similarity directly. In this respect, MDS differs from principal component analysis, factor analysis and correspondence analysis (Croft and Poole 2008: 13–14). The latter methods all involve *eigenanalysis*.[5] Eigenanalysis takes the matrix of data and converts it to another matrix of the same dimensionality as the original, such that (a) each dimension is uncorrelated with every other dimension and (b) the first dimension accounts for the most variance in the data, the second for the next

resentation, they will only be connected with very low-weighted edges. In contrast, a Euclidean space model always places all meanings in the spatial representation of similarity. That is, all meanings have a defined degree of similarity to all other meanings, that is, their Euclidean distance in the space, no matter what is the goodness of fit of the model.

5 Cysouw (2010) and Wälchli and Cysouw (2012) describe their representation of co-expression universals as "MDS", and indeed the R package they used calls it "MDS". However, the R package outputs an eigenanalysis (Wälchli and Cysouw 2012: 683–689); the output is not a distance model.

most variance, and so on. Typically, two dimensions of an eigenanalysis are displayed at a time, such as the first vs. the second dimensions, or the second vs. the third dimensions, due to the constraints of two-dimensional pages or screens.

An MDS spatial model, say a two-dimensional spatial model, represents all of the variance in the data in the reduced number of dimensions. A display of two dimensions of an eigenanalysis represents only a subset of the variance, namely the variance captured in the two dimensions displayed. Hence only an MDS spatial model is a true Euclidean spatial representation of the variation in the data. A two-dimensional display of an eigenanalysis is only a visual representation of the variance in the two principal components displayed – usually the first two principal components, which capture the largest amount of the variance.

Hence in an MDS spatial model, all distances are interpretable. The analysis is therefore invariant under translation and rotation. This means that one is allowed to interpret an MDS spatial model using horseshoes or circles in the space; or if a two-dimensional interpretation is best, such as tense vs. aspect in the analysis of Dahl's questionnaire data, the two dimensions do not have to correspond to the x and y axes of the display.

In contrast, each dimension of an eigenanalysis represents only the proportion of variance captured by the factor represented in that dimension in the display of two dimensions (Croft and Poole 2008: 14). Hence in an eigenanalysis, each dimension must be interpreted separately and independently:

> It is customary to summarize the row and column coordinates in a single plot. However, it is important to remember that in such plots, you can only interpret the distances between row points, and the distances between column points, but not the distances between row points and column points.
> (http://www.statsoft.com/Textbook/Correspondence-Analysis/, accessed 7 June 2018)

This is another point that is sometimes misunderstood.

7 Conclusions

This article has emphasized the similarity between the graph structure representation and the Euclidean space representation of typological universals of co-expression. Both the graph structure and the Euclidean space representations are useful and legitimate visualizations of similarity for a complex and variable dataset. Both can be used to analyze either exemplars or more general semantic categories that are co-expressed in the cross-linguistic data. Both representational models can now handle large, complex datasets using algorithms that can auto-

matically generate the representation. Both models have goodness of fit statistics that can be used to come up with a tighter set of co-expression universals from which contentful semantic explanations can be inferred.

Graph structure and Euclidean space representations differ in their utility and interpretation. The graph structure representation is more useful when there is a small number of nodes (concepts) being compared. The result is easily visualized, and Regier et al. (2013) provide a quantitative method that reliably constructs the minimally connected co-expression graph for the data. For more complex data sets, with a larger number of concepts, the Euclidean space representation derived with MDS by unfolding is likely to provide a better visualization of the generalizations in the cross-linguistic data.

One interesting difference between the semantic map model and the MDS model is that the graph structure semantic map model represents a discrete underlying conceptual space, while the Euclidean spatial model represents a continuous underlying conceptual space (Croft and Poole 2008: 20–22). These two mathematical models suggest different approaches to semantic analysis. The graph structure model is more amenable to a discrete conceptual space, where concepts are defined by sets of discrete semantic components, features or properties, of the sort suggested by Haspelmath for the conceptual structure of the indefinite pronoun space (Haspelmath 1997: 119–122). It is perhaps this assumption about discreteness that has led users of graph structure representations to look for discrete regions of the network via clustering or community-finding algorithms (e. g., List et al. 2013: 250; Jackson et al. 2019; Georgakopoulos and Polis 2021).

A continuous underlying conceptual space representation is most suited for usage-based or exemplar theories of language (Bybee 2010; Croft 2010c; Wälchli and Cysouw 2012: 674–676), where every instance of use or every semantic situation type is distinct but mapped onto a continuous space of semantic variation, or in phonology, where every sound production is distinct but mapped onto a continuous phonetic space. A graph structure model with a finite set of nodes would have to posit a very large number of such nodes in order to approximate the near-continuous conceptual space proposed by usage-based/exemplar theories of language.

Finally, representations of typological universals of co-expression cannot be simply equated with explanations in terms of the structure of the conceptual space. The data collection methods and the mathematical representations are tools to reach an explanation. Assuming good data from a well-stratified typological sample, the mathematical representations – derived algorithmically, and using fitness statistics to produce the best balance between informativeness and goodness of fit – will produce the most important semantic clusters or dimensions governing co-expression. Those semantic clusters or dimensions can then

be used to construct a semantically coherent conceptual space, and anomalies of co-expression in the semantically constructed conceptual space can be examined and accounted for by diachronic semantic and other explanations.

Acknowledgment: I would like to thank Stéphane Polis, Athanasios Georgakopoulos, Terry Regier, Johan van der Auwera and an anonymous reviewer for their comments on an earlier version of this article. All remaining errors are my own.

Appendix. Online resources

The unfolding algorithm for MDS analysis of linguistic data was implemented by Jason Timm, and is available at: https://github.com/jaytimm/MDS_for_Linguists.

The algorithm by Regier et al. (2013) for automatic generation of graph structure semantic map analysis is available at: http://lclab.berkeley.edu/regier/semantic-maps/.

CLICS is available at: https://clics.clld.org.

References

Bakker, Ryan & Keith T. Poole. 2013. Bayesian metric multidimensional scaling. *Political Analysis* 21. 125–140.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Cristofaro, Sonia. 2010. Semantic maps and mental representation. *Linguistic Discovery* 8(1). 35–52.

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

Croft, William. 2003. *Typology and universals*, 2nd edition. Cambridge: Cambridge University Press.

Croft, William. 2010a. Relativity, linguistic variation and language universals. *CogniTextes* 4. https://doi.org/10.4000/cognitextes.303.

Croft, William. 2010b. What do semantic maps tell us? Comment on 'Semantic maps and mental representation' by Sonia Cristofaro. *Linguistic Discovery* 8. 53–60.

Croft, William. 2010c. The origins of grammaticalization in the verbalization of experience. *Linguistics* 48. 1–48.

Croft, William. 2012. *Verbs: Aspect and causal structure*. Oxford: Oxford University Press.

Croft, William. To appear. *Morphosyntax: Constructions of the world's languages*. Cambridge: Cambridge University Press.

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

Cysouw, Michael. 2007. Building semantic maps: The case of person marking. In Matti Miestamo & Bernhard Wächli (eds.), *New challenges in typology: Broadening the horizons and redefining the foundations*, 225–247. Berlin & New York: Mouton De Gruyter.
Cysouw, Michael. 2010. Semantic maps as metrics on meaning. *Linguistic Discovery* 8(1). 70–95.
Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Basil Blackwell.
García Macías, José Hugo. 2016. From the unexpected to the unbelievable: Thetics, miratives and exclamatives in conceptual space. Ph.D. dissertation, University of New Mexico.
Georgakopoulos, Thanasis & Stéphane Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass* 12(2). e12270. https://doi.org/10.1111/lnc3.12270.
Georgakopoulos, Thanasis & Stéphane Polis. 2021. Lexical diachronic semantic maps. The diachrony of time-related lexemes. *Journal of Historical Linguistics* 11(3). 367–420.
Georgakopoulos, Thanasis, Eitan Grossman, Dmitry Nikolaev & Stéphane Polis. 2021. Universal and macro-areal patterns in the lexicon: A case-study in the perception-cognition domain. *Linguistic Typology*. https://doi.org/10.1515/lingty-2021-2088.
Greenberg, Joseph H. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of grammar*, 2nd edition, 73–113. Cambridge, MA: MIT Press.
Haiman, John. 1985. *Natural syntax: Iconicity and erosion*. Cambridge: Cambridge University Press.
Hartmann, Iren, Martin Haspelmath & Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language* 38. 463–484.
Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.
Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language, vol. 2*, 211–242. Mahwah, NJ: Lawrence Erlbaum Associates.
Jackson, Joshua Conrad, Joseph Watts, R. Henry Teague, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray & Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science* 366. 1517–1522.
Keenan, Edward L. & Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8. 63–99.
Keenan, Edward L. & Bernard Comrie. 1979. Data on the noun phrase accessibility hierarchy. *Language* 55. 333–351.
Levinson, Stephen C., Sérgio Meira & the Language and Cognition Group. 2003. 'Natural concepts' in the spatial topological domain—adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language* 79. 485–516.
List, Johann-Mattis, Anselm Terhalle & Matthias Urban. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. In Alexander Koller & Katrin Erk (eds.), *Proceedings of the 10th International Conference on Computational Semantics—Short Papers*, 347–353. Stroudsburg, PA: Association for Computational Linguistics.
Majid, Asifa, Nicholas Evans, Alice Gaby & Stephen C. Levinson. 2011. The grammar of exchange: A comparative study of reciprocal constructions across languages. *Frontiers in Psychology* 2. https://doi.org/10.3389/fpsyg.2011.00034.
Malchukov, Andrej L. 2010. Analyzing semantic maps: A multifactorial approach. *Linguistic*

Discovery 8(1). 176–198.
Maslova, Elena. 2003. A case for implicational universals. *Linguistic Typology* 7. 101–108.
Poole, Keith T. 2000. Non-parametric unfolding of binary choice data. *Political Analysis* 8(3). 211–237.
Regier, Terry, Naveen Khetarpal & Asifa Majid. 2013. Inferring semantic maps. *Linguistic Typology* 17. 89–105.
Rogers, Phillip. 2016. Illustrating the prototype structures of parts of speech: a multidimensional scaling analysis. MA thesis, University of New Mexico.
Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7(13). https://doi.org/10.1038/s41597-019-0341-x.
Sasse, Hans-Jürgen. 1987. The thetic-categorical distinction revisited. *Linguistics* 25. 511–580.
Stassen, Leon. 1997. *Intransitive predication*. Oxford: Oxford University Press.
Stassen, Leon. 2009. *Predicative possession*. Oxford: Oxford University Press.
van der Auwera, Johan. 2013. Semantic maps, for synchrony and diachrony. In Anna Giacolone Ramat, Caterina Mauri & Piera Molinelli (eds.), *Synchrony and diachrony: A dynamic interface*, 153–176. Amsterdam: Benjamins.
van der Auwera, Johan & Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2. 79–124.
van der Auwera, Johan & Lauren Van Alsenoy. 2011. Mapping indefinites: Towards a Neo-Aristotelian map. *Selected Papers from the 19th International Symposium on Theoretical and Applied Linguistics* 19. 1–14. https://doi.org/10.26262/istal.v19i0.5475.
van der Auwera, Johan, Petar Kehayov & Alice Vittrant. 2009. Acquisitive modals. In Lotte Hogeweg, Helen de Hoop & Andrej Malchukov (eds.), *Cross-linguistic semantics of tense, aspect and modality*, 271–302. Amsterdam: Benjamins.
Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50. 671–710.
Youn, Hyejin, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft & Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences* 113(7). 1766–1771. https://doi.org/10.1073/pnas.1520752113.