Inferring semantic maps

Terry Regier
Naveen Khetarpal
Asifa Majid

Abstract: Semantic maps are a means of representing universal structure underlying cross-language semantic variation. However, no algorithm has existed for inferring a graph-based semantic map from data. Here, we note that this open problem is formally identical to the known problem of inferring a social network from disease outbreaks. From this identity it follows that semantic map inference is computationally intractable, but that an efficient approximation algorithm for it exists. We demonstrate that this algorithm produces sensible semantic maps from two existing bodies of data. We conclude that universal semantic graph structure can be automatically approximated from cross-language semantic data.

Keywords: semantic universals, semantic maps, language and cognition

1. Introduction

Languages vary in their semantic categories – that is, in the range of semantic functions or uses picked out by their linguistic forms. However, many possible semantic categories are not attested, and similar categories often appear in unrelated languages. This pattern of constrained variation suggests a universal conceptual basis underlying the variation, such that different languages provide different snapshots of the same conceptual terrain. A SEMANTIC MAP is a means of capturing this idea, representing both presumed universal structure and language-specific partitionings of that structure.

A semantic map often takes the form of a discrete graph structure (e.g. Bybee, Perkins, & Pagliuca, 1994; van der Auwera & Plungian, 1998; Haspelmath, 1997). More recently semantic maps based on continuous representations have also been proposed (e.g. Croft & Poole, 2008; Cysouw, 2001; Cysouw, 2007; Levinson, Meira, & the language and cognition group, 2003; Majid, Boster, & Bowerman, 2008). In both traditions, the inferred underlying structure is sometimes interpreted as capturing the conceptual similarity between different semantic functions (e.g. Croft, 2003; Croft & Poole, 2008); in other work, no such attribution is made, and a semantic map is viewed simply as a compact description of attested variation, leaving open the possibility that the structure of the map may reflect extra-cognitive, such as diachronic or communicative, factors (e.g. Bybee, Perkins, & Pagliuca, 1994; Cristofaro, 2010). A carefully neutral statement of the purpose of a semantic map is that it attempts to "visually represent cross-linguistic regularity in semantic structure" (Cysouw, Haspelmath, & Malchukov, 2010: 1). In this paper, we use the term *semantic map* to refer specifically to graph-based maps, and we do not assume that the structure of the graph must necessarily accurately reflect cognitive reality – although we agree with Croft (2010) that it is likely to often do so.

Formally, a (graph-based) semantic map is a graph in which vertices (nodes) represent semantic functions or uses, and edges (links) connect closely related semantic functions. For a given semantic map, the semantic functions and the connections between them are assumed to be universal. The meaning of a given linguistic form is then represented as a language-specific grouping of vertices into a CONNECTED REGION of the universal graph.

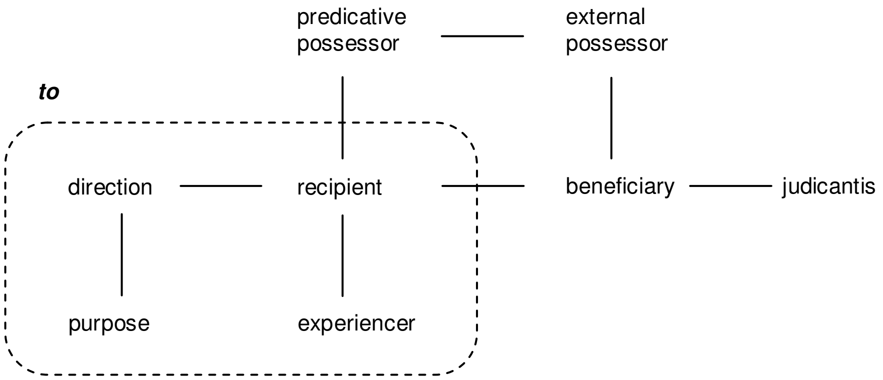

Figure 1. A semantic map of typical dative functions, with the semantic range of English *to* shown in dotted outline. French *à* is similar to English *to*, but excludes PURPOSE and includes PREDICATIVE POSSESSOR. From Haspelmath (2003: 213).

An example is shown in Figure 1. This semantic map, from Haspelmath (2003: 213), shows a set of typical semantic functions of the dative, and also shows the semantic range of the English word *to* as a connected subset of this universal graph. This *to* subset comprises the functions DIRECTION (e.g. She went *to* Philadelphia), RECIPIENT (e.g. He gave the book *to* his sister), EXPERIENCER (e.g. That seems loud *to* me), and PURPOSE (e.g. I did it *to* see what would happen). French *à* occupies an overlapping but distinct connected subset, and forms from other languages occupy yet other connected subsets. The SEMANTIC MAP CONNECTIVITY HYPOTHESIS (Croft, 2003: 134) is the proposal that language-specific categories will always pick out connected subsets of the graph. For example, given the semantic map in Figure 1, this hypothesis predicts that any linguistic form that expresses both RECIPIENT and PURPOSE will also express DIRECTION, since any connected region containing both RECIPIENT and PURPOSE must also include DIRECTION.

This hypothesis captures the widely-shared intuition that linguistic categories denote connected regions of conceptual or perceptual space: cf. Nerlove and Romney's (1967) observation that languages tend to avoid disjunctively defined kinship categories, and Roberson's (2005) notion of 'grouping by similarity' in color naming. Once a semantic map has been constructed to fit a body of cross-language data, the expectation is that new categories from as-yet-unexamined languages will also pick out connected subgraphs – possibly novel connected subgraphs. A semantic map thus compactly represents what patterns of variation one may and may not expect to find in a given semantic domain, and the underlying graph has been taken to represent "a common human cognitive heritage" (Croft, 2003: 139). Semantic maps have been widely used to represent cross-language semantic variation over a presumably universal base; for recent reviews see Haspelmath (2003) and Cysouw, Haspelmath, and Malchukov (2010) plus other papers in the same volume.

The task of constructing a semantic map in graph form from cross-language data is generally done by hand, and the task can be time-consuming with moderate to large-sized datasets. It would therefore be useful to automate this process; however the computational problem of inferring such a universal semantic map from cross-language data has not been formally addressed. Croft and Poole (2008) conjectured that this problem may be computationally intractable, and they considered this potential intractability to be a shortcoming of graph-based semantic maps as a representational tool in semantic typology. In contrast, a continuous map may be straightforwardly inferred from data using well-known computational techniques such as multidimensional scaling, and this fact has been held to be an advantage of continuous over graph-based representations for semantic maps (Croft & Poole, 2008; Cysouw, 2001; Wälchli, 2010). Here, we address the SEMANTIC MAP INFERENCE problem in formal terms, in the previously unexplored case of graph-based semantic maps.

In what follows, we first note that the semantic map inference problem is formally identical to another problem that superficially appears unrelated: inferring a social network from outbreaks of disease in a population. Angluin, Aspnes, and Reyzin (2010) have recently shown that this social network inference problem is computationally intractable, but that an efficient algorithm exists that approximates the optimal solution nearly as well as is theoretically possible; it follows that both the computational intractability and the applicability of the approximation algorithm hold of semantic map inference. We then apply this algorithm to the cross-language data of Haspelmath (1997) on indefinite pronouns, and of Levinson, Meira, and the Language and Cognition Group (2003) on spatial categories, in both cases yielding sensible and useful results. We conclude that presumptively universal structure consistent with cross-language semantic data can be straightforwardly inferred, that the issue of computational intractability—while real—need not deter researchers, and that formalization of problems in semantic typology can highlight useful connections to structurally related problems elsewhere.

2. The semantic map inference problem

The semantic map inference problem can be stated informally as follows. We are given a set of semantic functions or uses within a particular semantic range (e.g. RECIPIENT, PURPOSE, DIRECTION, etc. from the range of the dative, as in Figure 1). We are also given a set of groupings of these functions into semantic categories from various languages; each such grouping picks out the semantic functions that may be expressed by a given linguistic form (e.g. the functions of English *to* shown in dotted outline in Figure 1). We assume that each such category picks out a connected region of an underlying universal network of semantic functions, but we are not given the connections of that network. Instead, we wish to INFER the set of connections between semantic functions that best explains the observed semantic categories.

This problem can be formalized as follows, illustrated in Figure 2. Given a set V of vertices (representing semantic functions), and a set of constraints $S_i \subseteq V$ (representing a set of language-specific groupings of these functions into categories), we wish to find the minimum set of edges E between the vertices of V such that each S_i picks out a connected subgraph of the graph $G=(V,E)$. By asking for the minimum set E we avoid trivial and uninformative solutions such as those in which all vertices are connected. Moreover, because edges are inferred rather than directly observed, the existence of each edge must be assumed; this means that by minimizing the number of edges, we minimize the number of assumptions made, and thus privilege parsimonious solutions to the problem.

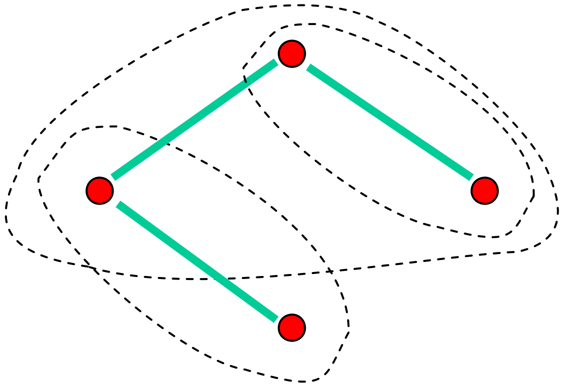

Figure 2. Formalization of the semantic map inference problem. We are given a set of semantic functions (vertices *V*, shown as small circles), and groupings of these functions into language-specific categories (constraints $S_i \subseteq V$, each shown by a dashed outline). We seek the minimum set of edges *E* (shown as links between vertices) such that each grouping picks out a connected region of the overall graph *G=(V,E)*.

Angluin et al. (2010) treated a formally identical problem. They wished to infer a social network from observations of disease outbreaks in a population. Thus vertices *V* now represent people, and each constraint $S_i \subseteq V$ represents the subset of people observed to have been affected by a particular disease outbreak *i*. For example, a particular S_i might represent the set of people observed to have caught a cold last November. Angluin et al. (2010) assumed that disease is spread by social contact, and they represented social contact between two people as an edge between the corresponding two vertices. They wished to find the social network that could best account for the observed outbreaks – that is, the minimum set of edges *E* such that each constraint S_i picks out a connected subgraph of the overall social graph *G=(V,E)*. This social network inference problem is formally the same as the semantic map inference problem; therefore any formal results concerning one also apply to the other.[1] (See also Dahl (2001: 1469) for a different disease analogy concerning grammaticalization.)

Some problems can be shown to be computationally intractable, in the sense that it is expected that there does not exist an efficient algorithm that will always find the optimal solution (Garey & Johnson, 1979). If a problem is computationally intractable in this sense, it is natural to abandon the search for an optimal solution and to ask instead whether an approximation to the optimal solution can be found efficiently. For some problems it can be shown that even this fallback goal of approximation is hard (e.g. Trevisan, 2004; Vazirani, 2001, ch. 29), meaning that there exists a value *r* such that no efficient algorithm can be expected to always approximate the optimal solution to within a factor of *r*. Angluin et al. (2010) showed that the social network inference problem is hard to approximate in this sense; therefore the same holds of the semantic map inference problem. This result confirms Croft and Poole's (2008) suspicion: the semantic map inference problem is indeed computationally intractable, and moreover is hard to approximate. However, this finding leaves open the possibility that an efficient algorithm may nonetheless produce approximations that are of high enough quality to be useful.

[1] Angluin et al. (2010) considered several variants of the social network inference problem. The specific variant to which we refer here is the one they label the offline uniform cost network inference problem; it corresponds to traditional graph-based semantic maps with unweighted edges. Other variants discussed by Angluin et al. (2010) are applicable to the suggestion (Cysouw, 2007: 233) that edges in semantic maps may usefully be weighted, to capture how often a given pair of semantic functions co-occurs.

3. The network inference algorithm

Angluin et al. (2010) presented an efficient algorithm for the social network inference problem and proved that it approximates the optimal solution nearly as closely as theoretically possible. Following the statement of the inference problem above, their algorithm is given a set *V* of vertices (which in the case of semantic map inference represent semantic functions), and a set of constraints $S_i \subseteq V$ (which in the case of semantic maps represent a set of language-specific groupings of these functions into categories). It begins with no edges *E* between the vertices. It then introduces edges one by one in order of their UTILITY (specified below), until each constraint S_i picks out a connected region of the overall graph.

Informally, the utility or usefulness of a proposed edge is the extent to which it contributes to the overall goal of the algorithm, namely a graph in which each constraint S_i picks out a connected region. For example, in Figure 2, it is visually clear that the already-inserted edge in the upper right portion of the graph (call it *e*) contributes to the connectedness of two constraints, whereas other already-inserted edges and other possible edges (not shown) each contribute to the connectedness of one constraint or no constraints. For this reason, beginning with no edges at all, *e* would have the highest utility and would be the first edge to be introduced.

This informal notion is captured formally by Angluin et al. (2010) by relying on the notion of a CONNECTED COMPONENT. A connected component of a graph is a maximal connected subgraph – that is, a connected subgraph to which no further vertices may be added without losing this connectedness. Consider again the graph in Figure 2. Prior to any edges having been inserted, the initial graph (consisting only of vertices) would have had 4 connected components, one corresponding to each vertex. The same graph but with only the above-identified edge *e* inserted, and no other edges, would have 3 connected components: one component consisting of *e* and the two vertices it connects, and one component for each of the two remaining vertices. Finally, the graph as it is shown contains just one connected component, because the graph as a whole is connected. With this by way of background, the Angluin et al. (2010) algorithm operates as follows.

Let s_i denote the subgraph of $G=(V,E)$ that is picked out (induced) by constraint S_i, and let ncc_i denote the number of connected components within s_i. When there are no edges connecting the vertices of s_i, ncc_i equals the number of vertices in s_i; this is its maximum possible value. When s_i is connected, ncc_i equals 1, its minimum possible value. In general, the lower the value of ncc_i, the closer constraint S_i is to being satisfied, i.e. the closer the subgraph induced by S_i is to being connected. Let *C* be an objective function defined as:

$$C = \sum_i (1 - ncc_i)$$

The algorithm begins with an empty edge set *E*. This yields a strongly negative value for *C* (except in the trivial case in which each constraint contains only one vertex). The algorithm then adds to *E* the edge that yields the greatest increase[2] in *C*. This steepest ascent step is repeated

[2] There may be instances in which more than one edge yields the same maximal increase in C. In such circumstances, the choice between these possibilities is not specified by the algorithm statement given here, and our implementation chooses among these possibilities arbitrarily, by the order in which edges are considered.

until all constraints are satisfied, i.e. until $C = 0$. Python code implementing this algorithm may be found at http://linguistics.berkeley.edu/~regier/semantic-maps/

Because this is an approximation algorithm, it is not guaranteed to find the optimal solution to a given instance of the problem. For our purposes, the relevant question is whether the degree of approximation attained by this algorithm is adequate to produce high-quality semantic maps from cross-language data. We will consider a map to be high-quality if it is relatively parsimonious – i.e. if it accommodates the data using few edges – and we leave for future work the exploration of other criteria of success, e.g. correctly inferring independently known cognitive or diachronic connections. With this parsimony criterion in mind, we turn now to test the algorithm empirically, against two well-established bodies of such data.

4. Indefinite pronouns

Haspelmath (1997) examined the semantic uses of indefinite pronouns, such as *anybody*, *someone*, and semantically related forms in other languages, through a large-scale cross-language study. His primary database contained 140 semantic categories, each associated with a linguistic form, from a total of 40 languages. This database is presented in full in his 1997 book. Each category picked out some subset of the following 9 semantic functions, illustrated below with examples from Haspelmath (1997):

1. SPECIFIC, KNOWN TO SPEAKER: *Somebody* called while you were away: guess who!
2. SPECIFIC, UNKNOWN TO SPEAKER: I heard *something*, but I couldn't tell what kind of sound it was.
3. NON-SPECIFIC, IRREALIS: Please try *somewhere* else.
4. POLAR QUESTION: Did *anybody* tell you anything about it?
5. CONDITIONAL PROTASIS: If you see *anything*, tell me immediately.
6. STANDARD OF COMPARISON: In Freiburg, the weather is nicer than *anywhere* in Germany.
7. DIRECT NEGATION: *Nobody* knows the answer.
8. INDIRECT NEGATION: I don't think that *anybody* knows the answer.
9. FREE CHOICE: *Anybody* can solve this simple problem.

For example, the English form *someone* can serve the following 5 semantic functions: SPECIFIC KNOWN, SPECIFIC UNKNOWN, IRREALIS, QUESTION, and CONDITIONAL. Based on the cross-language database, Haspelmath (1997: 64) constructed the semantic map shown in Figure 3. Each of the categories in his 40-language database corresponds to a connected subgraph of this graph, and the expectation is that the same will hold for forms from languages not yet examined.

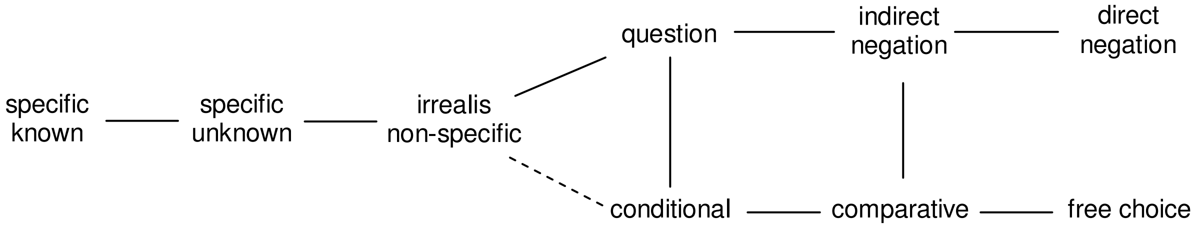

Figure 3. A semantic map for indefinite pronouns, adapted from Haspelmath (1997: 64). The dashed edge from IRREALIS NON-SPECIFIC to CONDITIONAL is included by Haspelmath, but not by Angluin et al.'s (2010) network inference algorithm.

Croft and Poole (2008) re-examined Haspelmath's (1997) 40-language database, and concluded that the edge from IRREALIS NON-SPECIFIC to CONDITIONAL is not necessary – that is, that the connectedness of each category in the database can be maintained without this edge. They took this finding to support their argument that "the best conceptual space is not easy to find by hand" (Croft & Poole, 2008: 6), and concluded that the absence of an automated method for inferring semantic maps from data is a potentially serious limitation.

We ran Angluin et al.'s (2010) algorithm on Haspelmath's 40-language database, which he kindly shared with us in electronic form, and obtained the semantic map suggested by Croft and Poole's observation – that is, the same map as Haspelmath's minus the one disputed edge. Thus this algorithm, and Croft and Poole, have found a simpler map than that provided by Haspelmath. Moreover, this simpler map is guaranteed by the algorithm to be sufficient to account for the 40-language sample. Whether this simpler map will also account for further data remains an open question. Haspelmath (1997: 64) states that his map was based both on the 40-language sample and on some data beyond it, so it is possible that the disputed edge is necessitated by data outside the sample. However, whatever the outcome of that question, the present study demonstrates that Angluin et al.'s (2010) algorithm produces output that is comparable in quality (parsimony) with an influential published semantic map, and thus establishes the usefulness of this algorithm as a means for inferring universal structure from cross-language data.

5. Spatial categories

Having tested the algorithm against a dataset that covers a small number of semantic functions or uses, we wished to further assess it using a dataset that covers a greater number. This would be very time-consuming to do by hand; it is presumably for this reason that most published semantic maps are small. We had two specific goals. The first was to determine whether the structure produced by the algorithm over this more complex domain was intuitively sensible. The second goal was to determine whether the inferred structure would accommodate data from a language other than those considered in building the map – that is, whether the structure inferred by the algorithm would generalize beyond the training set.

We conducted this test in the semantic domain of spatial relations. Spatial categories across languages show both universal tendencies and cross-language differences, as illustrated in Figure 4 and supported in greater detail by Bowerman (1996), Levinson et al. (2003), and Talmy (2000), among others. This mixture of universals and variation seems in principle capturable in terms of a semantic map – and indeed it has been captured in terms of continuous maps (e.g. Croft & Poole, 2008; Levinson et al., 2003). We sought to accommodate the same data using a large-scale automatically constructed graph-based map.

We relied on an existing dataset of cross-language spatial naming data, based on 71 pictures portraying simple spatial relations. These stimuli were originally designed by Bowerman and Pederson (1992; 1993); the scenes in Figure 4 are adapted from scenes in this set. Levinson et al. (2003) analyzed the spatial terms applied to these pictured spatial relations by speakers of 9 unrelated languages: Basque, Dutch, Ewe, Lao, Lavukaleve, Tiriyó, Trumai, Yélî-Dnye, and Yukatek. They describe the spatial naming data elicitation technique as follows: "Each picture has a designated FIGURE (or theme or trajector) colored yellow, and a GROUND object (or relatum or landmark), and the researcher uses the pictures to set up a verbal scenario as close as culturally possible to that depicted, and asks the consultant to answer a question of the form:

'Where is the [Figure]?' (given the sketched scenario)." (Levinson et al., 2003: 487). Levinson and Meira kindly shared with us the spatial naming data they had available, resulting from elicitation sessions with speakers of the above 9 languages, against the 71 scenes described above. We took these data as our dataset.

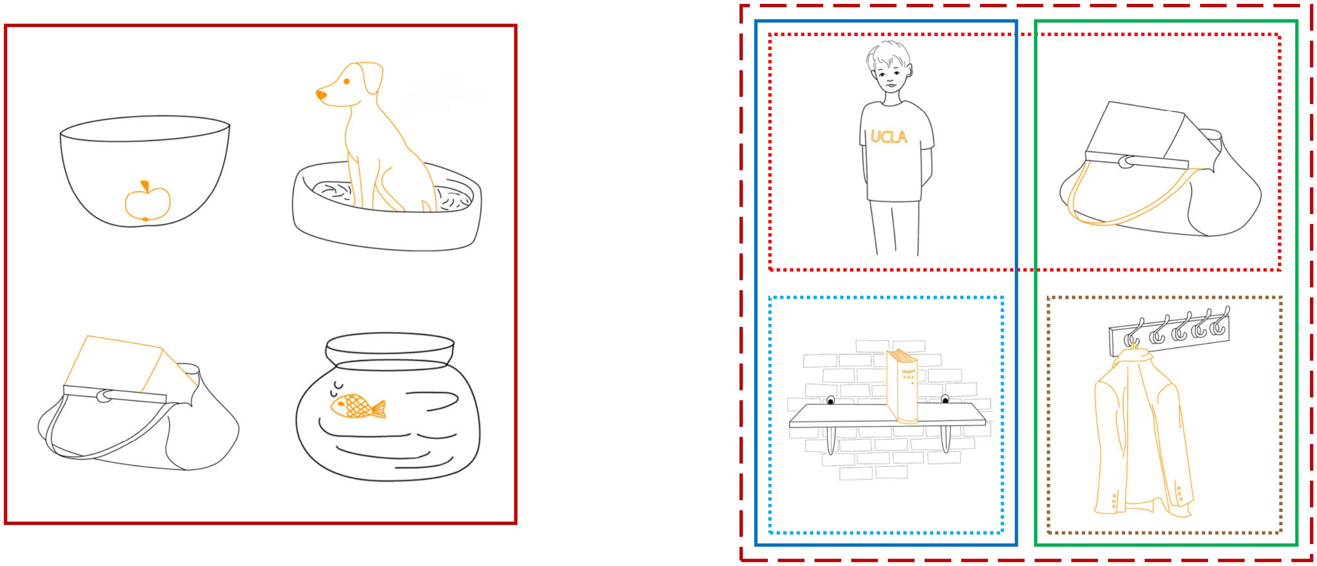

Figure 4. Universal tendencies and variation in spatial categorization. The 4 spatial relations in the left panel all fall in the same category in English (*in*), and also all fall in a single category in Dutch and in Yélî-Dnye. The 4 spatial relations in the right panel all fall in the same category in English (*on*; long dashed outline), but they are categorized differently in Dutch (solid outlines) and Yélî-Dnye (short dashed outlines). Based on the spatial dataset we treat in this paper.

Our treatment of the data followed theirs as closely as possible. They describe their data treatment as follows: "[E]ach language was treated on its own. An average of the consultants' responses was calculated: for the languages with many consultants … a picture was ascribed to a certain adposition when more than 50% of the consultants used it; for languages with four or five consultants, a picture was ascribed to a certain adposition if at least two of them used it; for the languages with three or fewer consultants, a picture was ascribed to a certain adposition if any of the consultants used it." (Levinson et al., 2003: 503) We followed this procedure for those 7 of the 9 languages for which data from individual speakers was currently available. However, for the remaining 2 languages, Ewe and Yukatek, data from individual speakers was not available, and thus we could not follow the above procedure. For these 2 languages only, we instead used summary data provided by Sérgio Meira.

A natural means of assessing a semantic map is to first construct the map based on data from one set of languages (which may be considered the training set), and to then see whether the resulting map also accommodates data from other languages (the test set). By the definition of the semantic map inference problem, the categories in the training set are guaranteed to pick out connected subgraphs of the resulting map; what is not known is whether the categories in the test set will as well. They should, to the extent that the inferred structure accurately reflects universal constraints on semantic variation.

We took the data from the 9 languages in the dataset to be our training set, and we took the spatial terms of English, applied to the same stimuli, to be our test set. The three authors, all native speakers of English, each independently named each of the scenes in English. A scene was assigned to an English spatial term when at least two of the three authors used the term to name that scene.

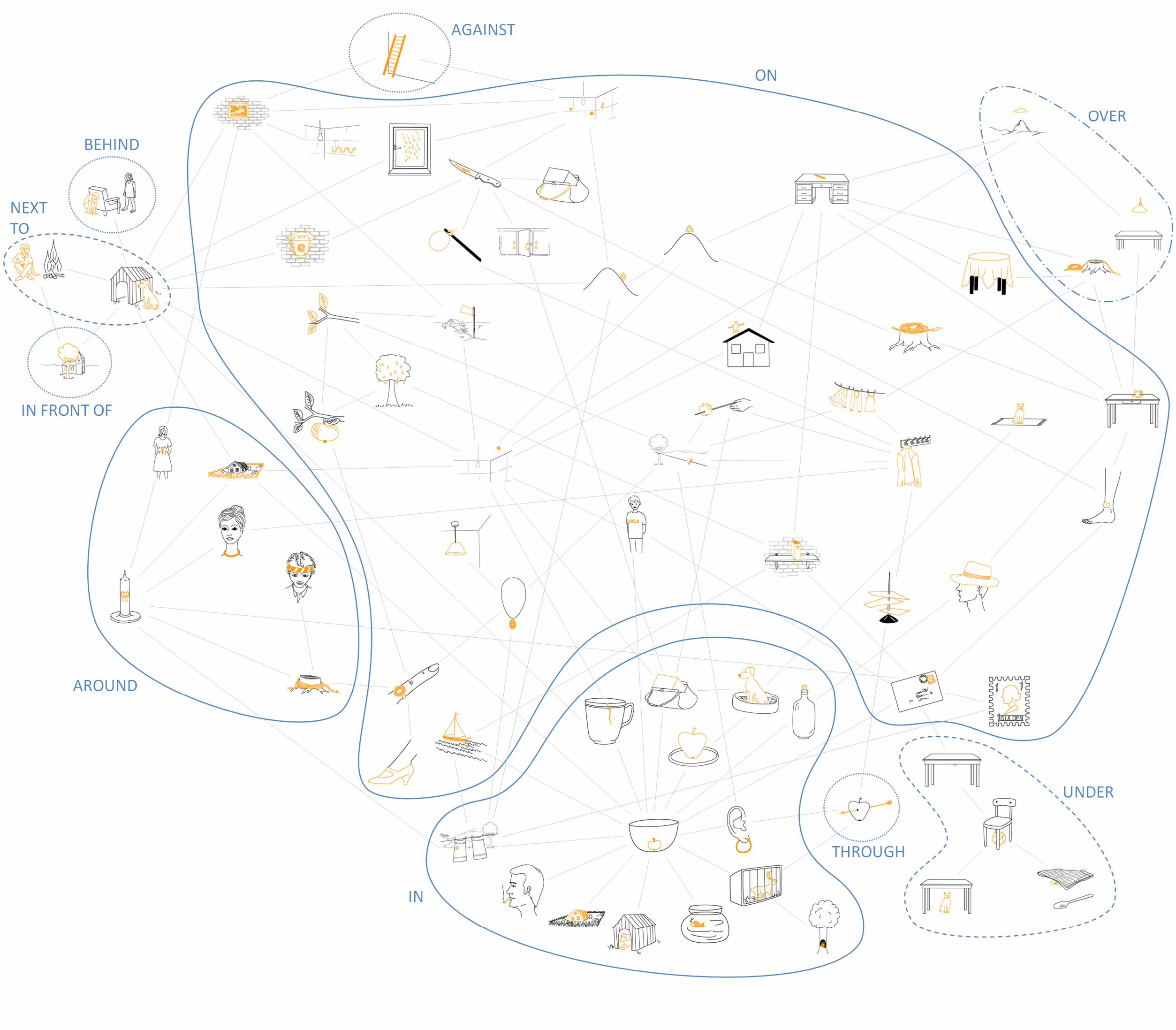

Figure 5. A semantic map of spatial meanings, obtained from Levinson et al.'s (2003) spatial language data. Spatial semantic categories of English are shown as outlined regions of this map. Dotted outline = singleton category; dashed outline = category present in the training data; solid outline = novel connected category; dotted-and-dashed outline = novel unconnected category. A higher-resolution version of this figure is available at http://linguistics.berkeley.edu/~regier/semantic-maps/

We obtained a semantic map from the training set via Angluin et al.'s (2010) network inference algorithm. The results are shown in Figure 5. Edges tend to connect closely conceptually related scenes; thus it appears that the inferred structure is intuitively sensible, at least on informal inspection. We then asked whether the spatial naming system of English was compatible with this map – that is, whether the spatial categories of English pick out connected subgraphs of the overall graph. Figure 5 shows that they do for all categories but one. There are four classes of English spatial category in this figure, distinguished by their outlines. The categories shown in light dotted outline (*against, behind, in front of, through*) contain only one scene each and are thus uninformative about connectivity. The categories shown in dashed outline (*next to, under*) correspond to connected subgraphs – but the same groupings were also present in the training set (associated with other, non-English, forms) and they are therefore necessarily connected in this map. The categories shown in solid outline (*around, in, on*) are informative: these categories are not present in the training data, and are nonetheless connected – thus they confirm a prediction implicitly made by the structure of the semantic map concerning what categories one may expect to find beyond the training data. Finally, the one category shown in dotted-and-dashed outline (*over*) is not present in the training data, and is not connected in this map – it thus violates the prediction that novel categories should conform to the induced structure.

How are we to evaluate these results? Unlike the case of indefinite pronouns discussed above, in the case of the spatial dataset there is no human-generated graph-based semantic map to which we may compare our results; they must be evaluated in their own terms. Strictly speaking, the model has failed to accommodate all the English data. At the same time, it has succeeded in accommodating almost all the data. What we need, rather than a categorical designation of success or failure, is a quantitative measure of degree of fit. There is no standard measure of degree of fit for graph-based semantic maps (Cysouw, 2007: 228), so we propose our own. Specifically, we use the objective function C described in section 3 above. That function reaches its maximum value of 0 when each semantic category in the data corresponds to a connected region of the network. The extent to which C is less than 0 measures how far a given semantic map is from fitting the data perfectly.

In the case of the semantic map shown in Figure 5, tested against the English data shown in that figure, $C = -1$. This is close to the ideal of 0, but that fact leaves important questions unresolved: Is it in any sense surprising or informative that the map fits the English data to the degree that it does? Would any system of categories of complexity comparable to English fit the semantic map as well as this? Or does English fit the structure of the semantic map significantly better than other systems of comparable complexity would?

We sought to answer these questions through a permutation test, as follows. We began with the semantic map shown in Figure 5, in which each scene is labeled with an English spatial term. We then considered hypothetical variants that retained the same network structure, the same number of English categories, and the same number of scenes per category – but randomly reassigned which English labels were assigned to which scenes.[3] Thus we consider a space of possible naming systems that are of complexity comparable to English, but that differ in the ways they partition the spatial semantic map. We sampled 10^5 (100,000) such hypothetical systems, without replacement, and measured C for each. We found that the value of C obtained from the actual English data shown in Figure 5 was higher than for any of these hypothetical systems (min=-46, max=-11). We conclude that the English system fits the structure of the network better than do hypothetical systems of comparable complexity. Importantly, if the semantic map had been very densely connected, rearrangements of the labels should not have

[3] The procedure we used to create each such hypothetical variant is as follows. Randomly select one of the English spatial terms – call the number of scenes associated with this term k. Then select k random scenes and group them into a category. Continue by selecting another English term and creating the next category from the set of as-yet-uncategorized scenes. Repeat this until all scenes are categorized and all English terms have been selected.

affected the degree of fit very much because most possible categories in such a map would be connected. Thus the present outcome suggests that the inferred semantic map of Figure 5 provides a description of the data that is sparse (parsimonious) enough, and thus constrained enough, that it accommodates attested data better than it does arbitrary hypothetical data.

The semantic map of Figure 5 is based on a small set of languages, against a larger set of stimuli than is common. The map's approximation to universal structure is presumably correspondingly loose – as is suggested by its imperfect fit to a novel language, English. A more complete test of these ideas will require a larger set of cross-linguistic data. Nonetheless, these results do show that the network inference algorithm can produce interpretable and intuitively reasonable semantic maps with a large number of vertices, that at least some of the predictions the resulting map makes about categories from new languages are supported, and that the resulting map is relatively parsimonious. These findings support the proposal (e.g. Croft & Poole, 2008; Levinson et al., 2003) that a universal representation may underlie the substantial cross-language variation in spatial semantic systems.

These results also raise a more general theoretical question, concerning the adequacy of connectedness as a constraint on semantic categories. The map in Figure 5 supports the categories in the training set, and most of those in the test set, as connected regions – but it also supports many other connected regions that seem implausible as semantic categories. For example, one may trace an elongated connected region that starts at one corner of the figure and extends in a chain to the opposite corner, picking out a series of connected scenes that each seem conceptually related to their immediate neighbors in the chain, but that do not hang together as a whole, and that exclude other conceptually related scenes. Categories do often pick out short chains of related meanings. For example, Bowerman and Pederson (1993) have identified an apparently universal sequencing or chain of spatial meanings (a subset of those meanings explored here), ranging from IN to ON, such that spatial terms from different languages pick out different subchains of the overall chain; this is effectively a semantic map in the form of a chain, with spatial terms picking out subchains. But these were relatively short chains of meaning, and it seems counterintuitive that a category would have the extremely high degree of elongation shown by the chain imagined above, in the context of Figure 5, without any coherent and reasonably compact core or central region. Thus, these results underscore the previously-noted fact that connectedness appears to be too loose a constraint on category shape (Croft, 2003: 138; Cysouw, 2001: 609), and that categories may tend to be more compact and coherent than is suggested by this constraint alone. This question mirrors a debate in the literature on color naming, over whether color terms pick out merely connected regions of perceptual color space, that might exhibit high degrees of chaining or elongation (Roberson, Davies, & Davidoff, 2000; Roberson, 2005) or regions that are both connected and compact (Jameson & D'Andrade, 1997). Although the question is implicit in the use of connectedness as a constraint in semantic maps generally, it becomes especially prominent given large maps such as that in Figure 5 – and the creation of such maps is facilitated by the availability of an algorithm for inferring such maps from data.

6. General discussion

We have seen that the problem of inferring presumptively universal structure from cross-language semantic data is formally identical to the problem of inferring a social network from disease outbreaks in a population. From this identity it follows that semantic map inference is computationally intractable, confirming an earlier conjecture to this effect. However it also follows that an existing approximation algorithm for social network inference may be applied to linguistic data, and we have seen that this algorithm yields sensible results when applied to two cross-language datasets of semantic categories.

Several questions are left open by these findings. It is unclear how well this algorithm, or any approximation algorithm that may be proposed to replace it, will perform on other datasets. It is also unclear which semantic domains, and which questions within these domains, are best approached using graph-based semantic maps, rather than another means of inferring the universal bases of semantic variation – for example, continuous representations such as those produced by multi-dimensional scaling and similar procedures (e.g. Cysouw, 2001; Croft & Poole, 2008; Levinson et al., 2003; Majid et al., 2008). Finally, the present results highlight the possibility that connectedness may be too loose a constraint on category shape, but they do not determine how best to address this shortcoming: whether it is preferable to supplement connectedness by further constraints (e.g. Croft, 2003: 138), to use weighted edges that reflect the frequency with which pairs of semantic functions co-occur (Cysouw, 2007: 233), or to pursue a different account altogether, such as the view that semantic systems across languages reflect the need for informative communication (e.g. Jameson & D'Andrade, 1997; Kemp & Regier, 2012; Regier, Kay, & Khetarpal, 2007). Settling these open questions will require further investigation.

Nonetheless, two broad conclusions can be drawn. First, high-quality (that is, relatively parsimonious) semantic maps can be efficiently inferred from cross-language data, and the question of computational tractability should therefore not be viewed as an obstacle to using them. Second, and more generally, these results suggest that the formalization of problems in semantic typology can lead to insight from structurally similar problems in unrelated domains.

Acknowledgements

We thank Martin Haspelmath for sharing his data, and Stephen Levinson and Sérgio Meira for sharing theirs. We also thank four anonymous reviewers for their very helpful comments and suggestions. This work was supported by NSF under grants SBE-0541957 and SBE-1041707, the Spatial Intelligence and Learning Center (SILC).

References

Angluin, D., Aspnes, J., & Reyzin, L. (2010). Inferring social networks from outbreaks. In Hutter, M. et al. (Eds.), *Algorithmic Learning Theory 2010, Lecture Notes in Computer Science, 6331* (pp. 104-118). Berlin: Springer.

van der Auwera, Johan & Plungian, Vladimir (1998). Modality's semantic map. *Linguistic Typology 2*, 79-124.

Bowerman, M. (1996). Learning how to structure space for language: A cross-linguistic perspective. In P. Bloom, M. Peterson, M. Garrett, & L. Nadel (Eds.) *Language and space* (pp. 385–436). Cambridge, MA: MIT Press.

Bowerman, M. & Pederson, E. (1992). Topological relations picture series. In S. C. Levinson (Ed.), *Space stimuli kit 1.2*: 51. Nijmegen: Max Planck Institute for Psycholinguistics.

Bowerman, M. & Pederson, E. (1993). Cross-linguistic studies of spatial semantic organization. In *Annual Report of the Max Planck Institute for Psycholinguistics* 1992 (pp. 53–56).

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world.* Chicago: University of Chicago Press.

Cristofaro, S. (2010). Semantic maps and mental representation. *Linguistic Discovery, 8*, 35-52.

Croft, W. (2003). *Typology and universals: Second edition*. Cambridge, UK: Cambridge University Press.

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

Croft, W. (2010). What do semantic maps tell us? *Linguistic Discovery, 8*, 53-60.

Cysouw, M. (2001). Review of Martin Haspelmath, *Indefinite Pronouns* (1997). *Journal of Linguistics 37*: 607-612.

Cysouw, M. (2007). Building semantic maps: The case of person marking. In M.Miestamo & B. Wälchli (eds.) *New Challenges in Typology* (pp. 225-247). Berlin: Mouton.

Cysouw, M., Haspelmath, M., & Malchukov, A. (2010). Introduction to the special issue "Semantic maps: Methods and applications". *Linguistic Discovery 8,* 1-3.

Dahl, Ö. (2001). Principles of areal typology. In M. Haspelmath, E. König, W. Oesterreicher, & W. Raible (Eds.), *Language typology and language universals, volume 2*. (pp. 1456-1470). Berlin: de Gruyter.

Garey, M. & Johnson, D. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: Freeman.

Haspelmath, M. (1997). *Indefinite pronouns*. Oxford: Oxford University Press.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), *The new psychology of language, vol.* 2 (pp. 211-242). Mahwah, NJ: Erlbaum.

Jameson, K. & D'Andrade, R. (1997). It's not really red, green, yellow and blue: An inquiry into perceptual color space. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295-319). Cambridge, UK: Cambridge University Press.

Kemp, C. & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science, 336*, 1049-1054.

Levinson, S., Meira, S., & 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.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition, 109*, 235-250.

Nerlove, S. & Romney, A. K. (1967). Sibling terminology and cross-sex behavior. *American Anthropologist, 69,* 179-187.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences, 104*, 1436-1441.

Roberson, D., Davies, I.R.L., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a Stone-age culture. *Journal of Experimental Psychology: General, 129*, 369–398.

Roberson, D. (2005). Color categories are culturally diverse in cognition as well as in language. *Cross-Cultural Research, 39,* 56-71.

Talmy, L. (2000). How language structures space. In L. Talmy (Ed.) *Toward a cognitive semantics, Volume 1* (pp. 177-254). Cambridge, MA: MIT Press.

Trevisan, L. (2004). Inapproximability of combinatorial optimization problems. Technical report TR04-065, Electronic Colloquium on Computational Complexity.

Vazirani, V. (2001). *Approximation algorithms*. New York: Springer.

Wälchli, B. (2010). Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery, 8*, 331-371.