Similarity semantics and building probabilistic semantic maps from parallel texts

Bernhard Wälchli

University of Bern

Länggassstr. 49

3000 Bern 9, Switzerland

waelchli at isw unibe ch

Abstract

This paper deals with statistical (non-implicational) semantic maps, built automatically using classical multidimensional scaling from a direct comparison of parallel text data (the Gospel according to Mark) in the domain of motion events (case/adpositions) in 153 languages from all continents in 190 parallel clauses. The practical objective of is to present one way (among other possible ways) how semantic maps can be built easily and fully automatically from large typological datasets (Section 3). Its methodological objective is to demonstrate that semantic maps can be built in various ways and that sampling of languages and small differences in the method chosen to build a semantic map can have a strong influence on the results (Section 4), which does not mean that semantic space is arbitrary, but rather that it is dynamic (having stretching and shrinking dimensions). The theoretical aim of the paper is to discuss similarity semantics, the implicit theoretical basis behind the semantic map approach, and to show that similarity semantics is not novel, but has a long-standing tradition in philosophy and psychology (Section 2).

1. Introduction

This paper illustrates building probabilistic semantic maps on the basis of an exemplar database of local phrase markers (adpositions and case) in motion events in a dataset of 190

contextually-embedded situations from a massively parallel text, the Gospel according to Mark (henceforth Mark), in 153 languages from all continents. *Massively parallel texts* (Cysouw and Wälchli 2007) are texts translated into many languages and Mark is one of few texts where a large amount of linguistic diversity in all contients can be covered. The idea underlying *probabilistic semantic maps* is to model general trends in the semantic organization of categories. The closer two situations are represented in a semantic map the more likely it is that they are represented by the same category in any language in the database (Wälchli and Cysouw forthc.). Instead of assuming abstract functional domains, concrete instantiations of particular functions are considered (*contextually embedded situations*) as they are determined by given contexts. Functional domains will emerge in the analysis as clusters of situations if there is evidence for them in the cross-linguistic dataset. Parallel texts allow for a direct cross-linguistic comparison of contextually embedded examples without previous abstraction of language-particular systems and without previous classification of semantic contexts. This makes it possible to compile large databases of cross-linguistically comparable examples in a large number of diverse languages at the cost of some idiomaticity due to translation. However, using translations is actually nothing else than the practical implementation of the abstract idea of translational equivalence, which is pervasive in functional linguistics.

Two contextually-embedded situations encoded by local phrase markers are exemplified in (1) and (2) from Wolof and Finnish with their English equivalent (Early Modern English of the King James Version). *Local phrase marker* is a cover term for adpositions (pre- and postpositions) and case. The term "local phrase" denotes here any nominal, adverbial or pronominal expression of the ground in motion events (semantic roles of goal, source and companion), be it marked by an adposition and/or case or be it unmarked. As is common in

typology, this is a functional domain rather than a formal concept. The local phrase markers are given in boldface in the examples and in the glosses.

(1) Wolof (Niger-Congo; Northern Atlantic) [Mark 1:29]

...génn na-ñu **ci** jàngu bi, ñu... dem

...exit PERF-3SG **PP.PROX** church the, 3PL go

ci kër Simoŋ ak Andare.

PP.PROXhouse Simon and Andrew

'...when they were come **out of** the synagogue, they entered **into** the house of Simon and Andrew.'

(2) Finnish (Uralic, Finnic) [Mark 1:29]

Synagoga-**sta** he men-i-vät suoraan Simon-in ja Andreaks-en koti-**in**

synagogue-**ELA** they go-PST-3PL straight Simon-GEN and Andreas-GEN house-**ILL**

'...when they were come **out of** the synagogue, they entered **into** the house of Simon and Andrew.'

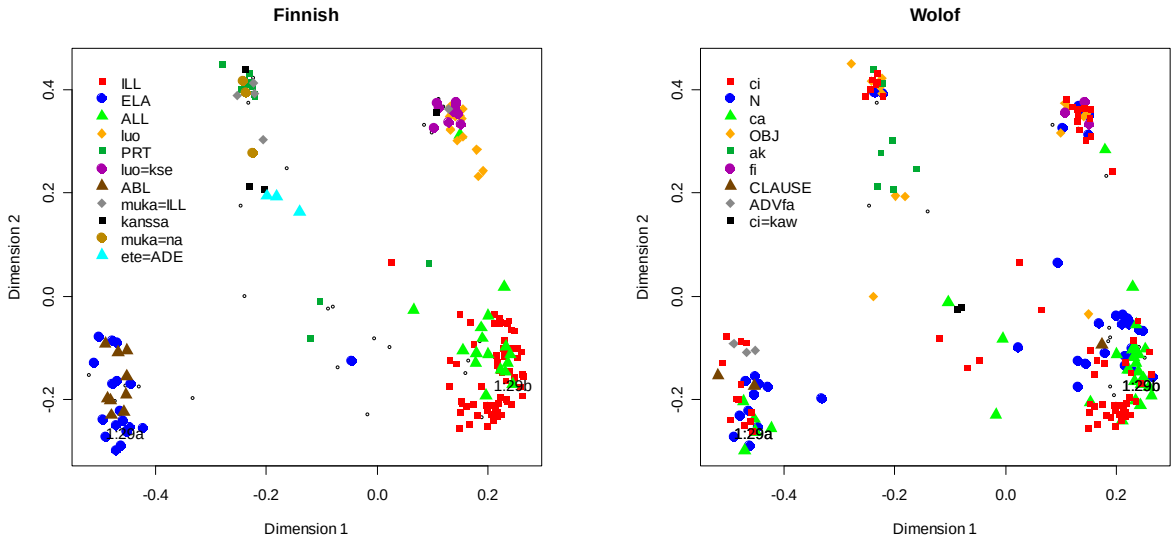
Finnish and Wolof have fundamentally different categorization patterns in local phrase markers. Finnish distinguishes the semantically opposite poles *source* and *goal* by means of case (Elative/Ablative for source vs. Illative/Allative for goal). In English, source (*out of, from*) and goal (*to, in[to], on[to]*) are distinguished by means of prepositions. However, Wolof does not distinguish source and goal in prepositions (and there is no case). The semantic categories expressed by Wolof prepositions are completely different: there is a distinction between proximal (*ci*) and distal (*ca*).

The semantic relationships depicted by a semantic map are often referred to as "*semantic space*" which is, of course, a metaphor that does not necessarily entail that there is a universal mental semantic space. Space in semantic maps is first of all *visualization*, which has two simultaneous, but partly conflicting aims: (a) a fully explicit (automatic) procedure to transform a part of the typological database into a graph with as little loss of detail (data reduction) as possible, and (b) a maximum of convenience of representation for the reader. What makes visualization difficult is that these two aims are sometimes in conflict.

Probabilistic semantic maps can be viewed as modeling the semantics of linguistic diversity and they do so to the extent that the sample (the underlying typological database) is representative of the population (the entire linguistic diversity). A general question addressed in many papers of this volume is whether semantic maps based on large typological datasets can model universal mental semantic space. This paper addresses that question from an empirical point of view. If semantic space is both mental and universal, it must be both comprehensive and robust. Robust means, different datasets (different samples of languages and of semantic functions) are assumed to yield highly similar maps representing the full range of semantic diversity encountered in natural languages. Comprehensive means, all semantic categories encountered in the database must be well-represented. It will be shown that the semantic map of local phrase markers (adposition and case) is neither robust nor comprehensive. Rather than reflecting the full range of cross-linguistic semantic diversity, semantic maps are a tool for identifying the fundamental tendencies in the data. Rather than yielding a single stable semantic map for all languages and all domains, semantic maps are dynamic assuming different shapes of constellations depending on the languages and functions sampled. This is consistent with the dynamicity of psychological similarity based on perception of situations discussed in Section 2.

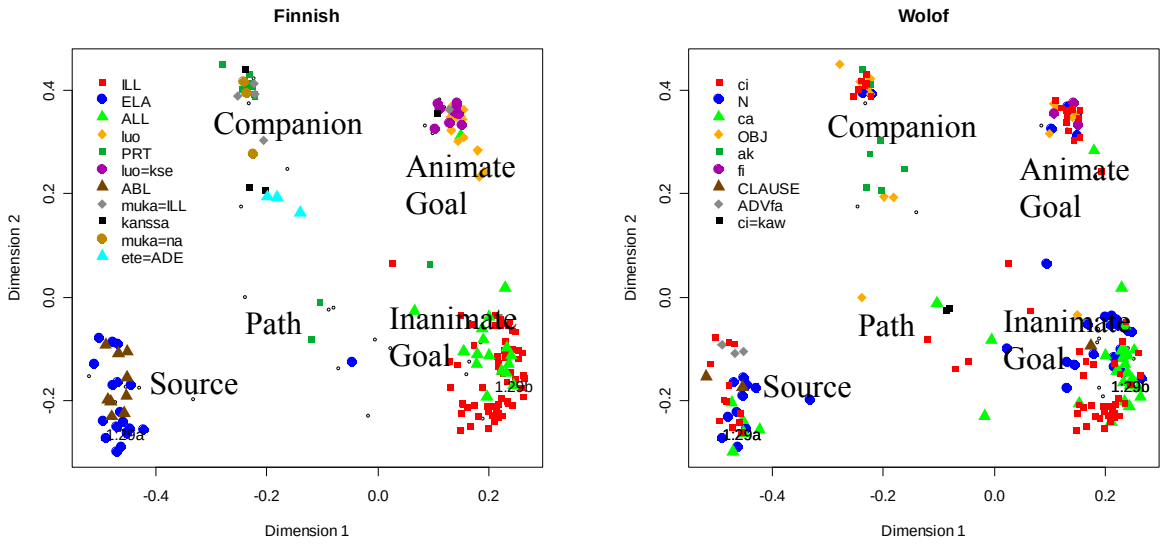
Exemplar-based semantic maps from parallel texts have the advantage that a large number of examples can be visualized in exactly the same configuration across all languages of the sample. The emerging configuration is such that situations are the closer to each other the more languages express them by the same form, which is done by visualizing a distance matrix by multidimensional scaling (MDS) (see Section 3). Figure 1 exemplifies the semantic map discussed in more detail in Section 3 for Finnish and Wolof. Each dot represents one of the 190 contextually-embedded situations in Mark and the positions of the two situations contained in examples (1) and (2) are indicated. The configuration of the dots represents the similarity relationships of all categories in all languages of the database. The symbol used to represent a category is determined by the category of the particular language displayed on the map, the category labels are given in a legend ordered according to their frequency of occurrence in the database. Thus, ILL (Illative) is the most frequent local phase marker in Finnish in the examples considered.

Figure 1: An exemplar-based semantic map of local phrase markers for Finnish and Wolof



It may now be objected that I have completely forgotten about aim (b) of visualization: a maximum of convenience of representation. As pointed out above, the two aims of visualization are difficult to reconcile. An important difference to classical semantic maps is that probabilistic semantic maps are not schematic. The configuration of the 190 dots is calculated automatically and the emerging clusters and the dimensions do not have any semantic labels. The clusters can now be interpreted by considering the semantic similarity of the situations clustering. It turns out that the first dimension (x-axis) distinguishes between source (negative pole, left, containing situation 1:29a in [1]) and goal (positive pole, right, containing situation 1:29b in [1])), while the second dimension (y axis) is sensitive to animacy, which is why animate goal is placed on top right and companion (mainly ground of "follow") on top middle. Figure 2 repeats Figure 1 with the major clusters labeled for convenience.

Figure 2: An exemplar-based semantic map of local phrase markers with auxiliary labels



Both Figures 1 and 2 and examples (1) and (2) serve to describe linguistic patterns by means of examples. However, Figure 1 does not only display two but 190 exemplar situations

and there is a database of 190·153 examples behind the configuration. This allows us to see that Finnish and Wolof have fundamentally different patterns of categorization in local phrase markers. The source-goal distinction in Finnish (Elative/Ablative vs. Illative/Allative) is a cross-linguistically very common distinction. Because it is supported by many other languages of the sample it emerges in the probabilistic semantic map as Dimension 1 even if completely absent in local phrase markers in Wolof and some other languages. However, the Finnish distinction between inner and outer local cases Illative/Elative vs. Allative/Ablative is much less clear-cut from a cross-linguistic perspective; and the dominant Wolof distinction in local phrase markers between proximal (*ci*) and distal (*ca*) is rare in local phrase markers. Other languages in the database do not support it, which is why these categories are not reflected as clusters on the semantic map (see Section 3 for discussion).

Unlike traditional implicational maps where the entities compared (the *analytic primitives*, see Cysouw 2007) are abstract functions with virtual translation equivalence, such as purpose, direction, and recipient, the basic entities to be displayed on the map considered here are contextually-embedded situations in a concrete text from real translations. It may be objected that the situations will not be identical if translation is not accurate and that translation always implies over- or underdetermination to a certain extent. Following this line of argument, the semantics of translational equivalents is hardly ever fully identical strictly speaking, but only very similar. This has the practical consequence for semantic maps that that the entities identified cross-linguistically should at least be more similar in meaning than the entities compared. In parallel texts situations are neatly determined by their textual embeddedness which makes it possible to include semantically more closely related situations among the analytical primitives to be compared.

Every semantic map has the properties *resolution* and *sharpness*. The degree of resolution is determined by the number of analytical primitives or “pixels”, to use a more common term

for indicating the resolution of pictures. Semantic distinctions which are more fine-grained than the analytical primitives chosen will never appear on a map. However, in order to yield a sharp map, the primitives (abstract domains or situations) identified across languages should be more similar than those compared within languages. Exemplar-based maps allow both for higher resolution (more situations considered) and for a higher degree of sharpness than maps representing abstract domains.

A main purpose of semantic maps is to make semantic analysis as empirical as possible by not making arbitrary ad hoc decisions. As pointed out in Haspelmath (2003:213), semantic maps are a method for approaching multifunctional patterns without implying "a commitment to a particular choice among monosemic and polysemic analyses." Semantic maps based on exemplar data go a step further. The analytical primitives are chosen such that they do not imply a commitment to a particular choice of abstract semantic domains. Rather semantic domains emerge in the map as clusters from exemplar situations if they are supported by the data. Unlike implicational maps which are claimed to display universal configurations (Haspelmath 2003:213), automatically built semantic maps from exemplar data are statistical.

The relationship between implicational and statistical semantic maps is the same as that between absolute and statistical universals. In recent typological research it has become clear that most universals are statistical rather than absolute (see the Konstanz Universals Archive). Restricting semantic map approaches to datasets which support implicational scales only would strongly limit the proportion of typological datasets that can be used to build semantic maps. It would also be against the spirit of the semantic map approach to incur as few commitments as possible before the analysis.

A similar line of argument is taken in Levinson and Meira (2003) and much other work in semantic typology from the Max Planck Institute in Nijmegen. As Levinson and Meira argue: "Generalizations about universal patterns must take into account that we are dealing with

much more diversity than the orthodox view suggests" (2003:513). The difference between the psycholinguistic approach of Levinson and Meira and the typological approach taken here is a difference in priority. The psycholinguists argue that "semantic data are not available without specially designed fieldwork" (Levinson and Meira 2003:492). They prioritize quality of data collection methods over large samples of languages. My point of view is that large samples are equally relevant which will be illustrated in Section 4 and that the loss due to distortions in translation is generally overestimated. An aspect of the semantics of local phrase markers which clearly suffers due to translation is absolute frames of reference (Levinson 2003). For instance, many languages of Oceania have adpositions or case indicating movement seaward, landward or parallel to the beach (see the discussion of Tobelo in Section 3). Such markers, even though not completely absent from Mark, are used more rarely than in original texts due to the difficulties of translation between different frames of reference and are therefore not given their representative weight in the database underlying the semantic maps built here.

Parallel texts, whatever their representativity for world-wide structural diversity, have some methodological advantages over potentially more reliable data. They allow cross-linguistic comparison on the level of contextually-embedded situations and they are more easily available. Probabilistic semantic maps provide a tool to do justice to the attested linguistic diversity while at the same time showing the main tendencies in the data material (a "typology without types"). One of their major advantages is that they are a tool for massive cross-linguistic comparison with little data reduction.

The practical objective of this paper is to present one way (among other possible ways) how semantic maps can be built easily and fully automatically from large typological datasets (Section 3). Its methodological objective is to demonstrate that semantic maps can be built in various ways and that sampling of languages and small differences in the method chosen to

build a semantic map can have a strong influence on the results (Section 4), which does not mean that semantic space is arbitrary, but rather that it is dynamic (having stretching and shrinking dimensions). The theoretical aim of the paper is to discuss similarity semantics, the implicit theoretical basis behind the semantic map approach, and to show that similarity semantics is not novel, but has a long-standing tradition in philosophy and psychology (Section 2).

2 The isomorphism hypothesis, similarity semantics, and exemplar semantics

The semantic map approach is heavily empirical. However, data and theory do not exclude each other. Typologists building semantic maps believe that constructing semantic models on the basis of large typological datasets is an indispensable approach to a better understanding of meaning which could not be reached by introspection in a single particular language or in a semantic metalanguage.

However, giving theoretical aspects high priority is indispensable because the theoretical basis of semantic maps, even though anything else but novel, is little known and does not play any major role in mainstream semantic theories, even though the semantic map approach has many predecessors in linguistics, philosophy, and psychology. A first step toward semantic maps is the rather trivial finding that categories are not identical cross-linguistically, but only similar.

An early philosophical pioneer of the semantic map method was Arthur Schopenhauer. He used overlapping circles to illustrate the non-congruence of concepts across different languages, illustrating the differences of "spheres of meaning" of words in different languages:

> „Nämlich sämmtliche Begriffe, welche zu bezeichnen die Worte der *einen* Sprache dasind, sind nicht grade durchweg dieselben, welche durch die Worte der *andern* Sprache bezeichnet werden; sondern sehr oft bloß ähnliche“[1] (Schopenhauer 1913:243).

However, he concentrates his discussion on adjectives such as *frappant*, *auffallend*, *speciosum* and abstract concepts, such as *amor*, *Liebe*, *pietà*, but severely overestimates the scope of identity in claiming that, for instance, *Baum*, *arbor*, *dendron* ‘tree’ have the same spheres of meaning (1913:243). The very same nominal domain—tree/wood/forest—served Louis Hjelmslev (1961:51-54 [1943:48-50]) to make a similar point about the non-congruence of linguistic categories, based on earlier work by de Saussure.

> “Each language lays down its own boundaries within the amorphous ‘thought-mass’ and stresses different factors in it in different arrangements, puts the centers of gravity in different places and gives them different emphases.” (Hjelmslev 1961:51-54 [1943:48-50])

Semantic maps are an indirect approach to the description of meaning. Similarity in meaning is accessed by way of formal identity (categories in particular languages) in a diverse set of languages. This approach is possible because there is a systematic exception to de Saussure’s arbitrariness of the sign. According to the *arbitraire du signe,* the relationship between form and meaning is accidental. However, the more similar two meanings, the more likely are they expressed by the same form in any language. This is known in the literature as Haiman’s *isomorphism hypothesis*:

---

[1] “All concepts for which the words of one language exist to denote them are not always the same as those which are denoted by the words of another language, but very often only similar concepts.” [translation BW]

(3) Haiman's isomorphism hypothesis

[A] "Different forms will always entail a difference in communicative function."

[B] "Conversely, recurrent identity of form between different grammatical categories will always reflect some perceived similarity in communicative function." (Haiman 1985:19)[2]

While Haiman's formulation of the isomorphism hypothesis is well suited as a basis for universal/implicational semantic maps it can only be applied to datasets where implicational relationships hold true without exceptions. Statistical/probabilistic semantic maps make a weaker claim and do not require that all situations of all categories in all languages cluster or are connected by lines (the Semantic Map Connectivity Hypothesis, Croft 2001:96). The isomorphism hypothesis must therefore be reformulated for probabilistic semantic maps as follows:

(4) Isomorphism hypothesis (weaker claim):

Given any two meanings and their corresponding forms in any particular language; the more similar the two meanings, the more likely that they are expressed by the same form in any language.

Put differently, categories have the property that they group similar rather than dissimilar exemplars together. This does not entail that similarity is a sufficient condition for

[2] The two parts of the isomorphism principle, notably [A], are also known by other names, e.g., Principle of Contrast (E. Clark 1993:69), *loi de répartition* (Bréal 1897/1913:26). According to Gilliéron (1919:9) formal "collision" of words in diachrony (two words becoming synonyms) provokes a fight in which one of the word is "killed".

categorization, but it is a necessary condition for categories. Categories consisting of membra disjecta cannot persist. This leads us to a view where similarity is not only required for describing the non-congruence of language-particular categories, but more generally for modelling any relationship between meanings.

Similarity semantics, as understood here, is a cover term for all approaches to semantics where similarity is considered to be a more basic notion than identity. The clearest representative of similarity semantics in philosophy is Fritz Mauthner. In Mauthner's view, similarity is the more fundamental notion than identity:

> „Absolute Gleichheit ist eine Abstraktion des mathematischen Denkens. In der Wirklichkeit gibt es nur Ähnlichkeit. Gleichheit ist starke Ähnlichkeit, ist ein relativer Begriff. Von der Schärfe der Sinnesorgane oder weiter des wissenschaftlichen Denkens, in letzter Instanz von der Aufmerksamkeit oder dem Interesse hängt es ab, wie weit z.B. eine Klassifikation getrieben wird...“[3] (Mauthner 1923:469).

For Mauthner, similarity is a necessary condition of language. Conceptualization is possible only because the senses are not sharp and humans therefore overestimate similarity. Identity semantics would be appropriate for omniscient subjects with exhaustive encyclopedic knowledge, such as Jorge Luis Borges' character *Funes el memorioso* who had a language where every individual thing had a name of its own (Borges 1944/2005:133; Borges mentions Mauthner explicitly as one of his sources of inspiration). While identity of

[3] "Absolute identity is an abstraction of mathematical thinking. Identity is strong similarity, is a relative notion. It depends on the sharpness of the senses and on the sharpness of scientific thinking, or put more generally, on degree of attentiveness and interest, how far, for example, a particular classification is driven." [translation BW]

two concepts can be established only if everything is known exhaustively about the two concepts, making judgments about the similarity of things is possible even for subjects who know very little:

> „Dabei möchte ich aber behaupten, daß diese bloße Ähnlichkeit, d. h. die wissenschaftliche oder mathematische Unvergleichlichkeit der Dinge erst unser Sprechen oder Denken möglich gemacht hat, daß also erst die Lücken unserer Vorstellungen, die Fehler unserer Sinneswerkzeuge unsere Sprache gebildet haben...Würde unser Gehirn von Natur auch nur annähernd so genau arbeiten wie Mikroskope, Präzisionsthermometer, Chronometer und andere menschliche Werkzeuge, würden wir von jedem Einzelding ein so scharfes Bild auffassen und im Gedächtnis behalten, dann wäre die begriffliche Sprache vielleicht unmöglich. Es wäre uns dann einfach versagt, den Begriff Anemone zu bilden; die einzelnen Anemonen wären einander zu unähnlich...die ganze Begriffsbildung der Sprache wäre nicht möglich, wenn wir nicht unter lauter lückenhaften Bildern umhertappten, eben wegen der Lückenhaftigkeit die Ähnlichkeit überschätzten und so aus der Not eine Tugend machten. Je weniger wir von etwas wissen, desto leichter werden wir von Ähnlichkeiten „frappiert“...So gebrauchen wir überhaupt Ähnlichkeitsbilder oder Worte umso leichter, je unwissender wir sind. So ist also die menschliche Sprache eine Folge davon, daß die menschlichen Sinne nicht scharf sind.“[4] (Mauthner 1923:437-438).

[4] “I would claim that it is similarity – that is, the scientific or mathematical incomparability of things – what has made possible that we speak and think. The gaps in our concepts, the shortcomings of our senses shape language...If our brain by nature worked only distantly as precisely as microscopes, precision thermometers, chronometers and other human tools, if we would retain from each particular thing such a sharp image in our mind, then a language based on concepts would perhaps be impossible. It would simply be impossible for us to form the concept anemone. The particular anemones would be too dissimilar...The whole conceptualization in

The meaning of a category can be approached in two different ways. It can be considered to denote an abstract concept or it can be considered to be a range of individual meanings of exemplars. Most current and ancient semantic theories assume that meaning denotes abstract concepts. However, exemplar semantics has an early philosophical predecessor in George Berkeley who rejected John Locke's notion of abstract ideas:

> "But it seems that a word becomes general by being made the sign, not of an abstract general idea, but of several particular ideas, any one it indifferently suggests to the mind" (Berkeley 1710/1998:94 [1710/1734:§11]).

Similarly, Ogden and Richards (1923/1966:99-101) reject the notion of concept ("conveniences in description, not necessities in the structure of things").

In a way similarity semantics, such as exposed by Mauthner, and exemplar semantics, such as exposed by Berkeley, is very disappointing from a philosophical point of view, because it leaves little room to a priori speculation. There are many ways in which two exemplars or situations can be considered similar or dissimilar, which is why similarity semantics is a fully empirical approach to meaning. This is why similarity has often been regarded as too unconstrained a notion, as being too flexible (Roberson 1999:2) or as Goodman (1972) puts it, similarity is "a pretender, an impostor, a quack" (437), "similarity is relative and variable, as undependable as indispensable", and "circumstances alter similarity"

---

language would not be possible, if we would not be groping in the dark under nothing but fragmentary images and if we would not – because of this fragmentarity – overestimate the similarity and so make a virtue of a vice. The less we know about something, the more we are astounded by similarities...This is why we use our similarity images or words the more easily the more ignorant we are. Therefore the human language is a consequence of the fact that the human senses are not sharp." [translation BW]

(445). The basic idea of similarity and exemplar semantics does not say anything more than that meaning is constituted by similarity relationships between exemplars rather than the meaning of entities and situations in isolation. However, the set of possible semantic links between two entities or situations is not a priori predictable as emphasized by Karl Otto Erdmann (1923).

Erdmann illustrates the unpredictability of semantic changes by examples where semantic change goes through accidental referents, such as French *grève* 'strike' deriving from French *grève* 'sandy beach of a river' by intermediation of the city hall square in Paris (formerly *Place de Grève*) where unemployed vagrants used to hang around (the example is attributed to K. Nyrop, Erdmann 1923:23). This semantic change of the category *grève*, presupposes familiarity with a particular referent with that name. In this case, the semantic change is very rare, probably unique, but if the particular referent with its accidental properties is familiar to all language communities, as in the case of "moon" > "month", a semantic change by way of a particular referent need not be rare.

While there are few works in modern philosophy and linguistics where the emphasis of semantic research is on the semantic links between items rather than the meaning of items in abstraction, the spirit of similarity semantics can be found implicitly and explicitly in many psychological and psycholinguistic works, such as, for instance, Mervis (1988):

> "Very young children, like adults, form object categories on the basis of similarity among exemplars. But judgments of similarity differ depending on the attributes to which a person attends. For example, consider the triplet robin, canary, lemon. Almost everyone would agree that the robin and the canary were the most similar pair. In this case, similarity is defined according to general form attributes. However, if the attribute 'yellow' were given sufficient weight, then the canary and the lemon would be the most

similar. Thus, in talking about categorization, the type of similarity which provides the basis for category assignments must be specified" (Mervis 1988:104-106).

Not incidentally, much psychological work on similarity is connected with color, the area where semantic space is not only an abstract postulate, but is directly accessible as a continuous perceptual space with measurable physical properties. However, there is a danger of overemphasizing perceptual similarity, as argued by Roberson et al. (1999). Roberson et al. (1999) discuss problems of invoking perceptual similarity to explain categorization. They report a series of experiments with a patient with language impairment with intact implicit judgments of categorization who fails in tasks tapping explicit categorization (naming, sorting colors into groups). His color and face freesort performance exhibit a marked adherence to pairwise similarity comparisons without revealing any effects of category boundaries. They conclude that perceptual similarity comparisons are insufficient to determine category membership without non-perceptual category-relevant information. Even if the implicit use of color and face categories is derived from an innately determined neural organization, the explicit use of these categories requires intact linguistic abilities.

Roberson et al. (1999:29) follow Goodman (1972) in claiming that similarity is a three-place relation, involving the two items to be compared and the respects relative to which the comparison is to be made.[5] Roberson et al.'s description of patient LEW's freesorting task, however, suggests that his similarity judgments lack the third place in the relation or have at least highly indeterminate respects relative to which comparison is made:

[5] However, Goodman also says that "[S]imilarity cannot be equated with, or measured in terms of, possession of common characteristics" (1972:443).

> "LEW looked for two stimuli that were the most perceptually similar. If satisfied that they met his criteria for grouping he placed them together, later using one of them to carry out the same procedure with another stimulus. With a large group of stimuli, this exercise took considerable time and on a number of occasions LEW declared himself dissatisfied with an emerging group and began to compare individual members to the members of other groups" (Roberson et al. 1999:9).

A more sophisticated model of similarity has been proposed by Nosofsky and Palmeri (1997:267), according to whom similarity between exemplars is a decreasing function of their distance in a multidimensional psychological space. Nosofsky and Palmeri (1997:267) trained subjects to learn two categories A and B represented by computer-generated color stimuli differing in brightness and saturation where both dimensions were relevant for classifying the objects. The subjects were asked to rate the similarity of pairs of stimuli by using a 10-point scale on which basis the arrangement of the stimuli in the individuals' psychological space could be modeled by a multidimensional scaling analysis. Nosofsky and Palmeri (1997:267) found that the response time in the categorization task correlates with the distance of a stimulus from the category boundary (the greater the distance of a stimulus from the exemplar-based boundary, the faster is the response time) and with familiarity of stimuli (familiar stimuli have shorter response time than unfamiliar given equal distance from category-boundaries). Nosofsky and Palmeri (1997) present an Exemplar-Based Random Walk Model (EBRW), which accurately predicts response times in categorization tasks not only for groups of test persons but for individuals. The same model can be used to predict old-new recognition judgments and response time of color-stimuli which varies depending on the degree of similarity of new stimuli with old stimuli (Nosofsky and Stanton 2006). In the EBRW model, when an item *i* is presented, it sets off a race among all exemplars stored in

the memory. The degree to which an exemplar *j* is activated is determined jointly by the exemplar's strength in memory and by its similarity to the presented item. Similarity is an exponential decay function of the distance *d* in the multidimensional similarity space (Shepard 1987). The exemplar that wins the race enters into the random walk. If it belongs to Category A then the random walk counter of that category is increased by unit, if it belongs to another category, the counter of Category A is decreased. The category whose category criterion is first reached is the response.

Nosofsky and Palmeri's (1997) EBRW model draws on Logan's (1988) Instance-Based Model of Automaticity, which is, however, identity-based. In Logan's model only exemplars that are identical to the presented item enter the race and the first retrieved exemplar initiates the action. In the EBRW model decisions are slower especially for objects difficult to discriminate which serves to predict response time accurately.

It seems to me that the evidence presented by Roberson et al. (1999) and Nosofsky and Palmeri (1997) are not in conflict. Nosofsky and Palmeri's (1997) notion of similarity cannot be abstracted from the notion of multidimensional psychological space. Furthermore, they do not discuss how categories emerge, but how category judgments are made. While Roberson et al. (1999) emphasize the importance of language and non-perceptual similarity, Nosofsky and Stanton (2006) emphasize that performance must be modeled at the individual-participant level. The structure of psychological space is not constant, but differs from individual to individual and across time. The distance between two exemplars in the space depends on attention weights for every dimension. Attending selectively to a dimension serves to stretch the space along that dimension and shrink the space along unattended dimensions. Put differently, according to this model semantic space is not universal, not even language-specific, but different for every individual and changing over time. For linguistic semantic maps this means that universal maps are only rough approximations. Semantic

similarity space does not only vary across languages but also across individuals and is dependent on the concrete exemplars individuals encounter and their order of presentation. Perception and categorization of exemplars interacts with the dynamic semantic space.

What is of particular importance for our purposes is the idea that semantic space, both if understood as psychological semantic space in individuals and averaged semantic space modeled in typological investigations, might be dynamic rather than static. While psychological semantic space changes as a consequence of different selected attention to different sets of exemplars, typological semantic space changes as a consequence of the sample of situations and languages sampled in the underlying database. Let us now first build a static typological semantic map based on an exemplar dataset (Section 3) and then explore how it changes if the sample of languages and situations is modified (Section 4).

3 Building a semantic map of local phrase markers from parallel text data

In this section we build a semantic map of local phrase markers (adposition and/or case) in 153 languages from all continents in 190 motion event clauses from translations of Mark. We will then explore in Section 4 how this map changes if the sample and the way of counting identity of categories are altered. Table 1 shows the processing chain in building the map and how it differs from traditional implicational maps.

Table 1: Processing chain in building semantic maps (following Cysouw 2007)

Approach	Analytical primitives	Set of empirical relations between every pair of primitives (distance matrix)	Graphical display, visualization tool
Implicational maps (Haspelmath 2003)	Abstract functions with virtual translation equivalence	Attested or unattested as combined into the meaning of a language-particular category	Connecting lines between related functions
Semantic maps from parallel texts (this paper)	Coding means in utterances in aligned parallel corpora	Hamming distance	Multidimensional scaling (MDS)

The languages of the sample are not languages properly but doculects. This term has been coined by Michael Cysouw, Jeffrey Good and Martin Haspelmath in 2006 to denote a variety of a language that has been described or otherwise documented. It is first mentioned in the published literature in Bowern (2008:8). Doculect is related to language as sample to population in statistics. In the ideal case, a doculect is fully representative of a language. However, for typological purposes and especially for the semantic map approach it is equally important that doculects are as directly comparable as possible (similar style and register and especially the same domains documented), and this is an advantage of Bible translations (Masica 1976:130, Wälchli 2007, but see also de Vries 2007). Wherever I use "language" below, this has to be understood in the sense of doculect.

Table 2: Sample (153 languages, wherever possible WALS names used):

Acholi, Adyghe, Ainu, Akan, Ambulas, Amuesha, Armenian (Classical), Avar, Aymara, Bambara, Bari, Basque, Batak (Toba), Breton, Bribri, Cakchiquel, Chamorro, Chiquito, Choctaw, Coptic, Cree (Plains), Creek, Dakota, Drehu, Efik, Enga, English, Estonian, Ewe, Fijian, Finnish, French, Garo, Gbeya Bossangoa, Georgian, Georgian (Classical), German (Bern), Greek (Classical), Greek (Modern), Guaraní, Haitian Creole, Hausa, Hawaiian, Hindi, Hmong Njua, Hopi, Hungarian, Icelandic, Igbo, Ijo (Nembe), Indonesian, Irish, Italian, Jabêm, Ju|'hoan, Kabba-Laka, Kabiyé, Kabyle, Kala Lagaw Ya, Kannada, Kâte, Khalkha, Khasi, Khmer, Khoekhoe, Kiwai, Komi-Zyrian, Korean, Koyra Chiini, Kriol (Fitzroy Crossing), Kuku-Yalanji, Kuna, Kunama, Kuot, Kurmanji, Latin, Lahu, Lak, Latvian, Lezgian, Lithuanian, Liv, Maltese, Mandarin, Maori, Mapudungun, Mari (Meadow), Marshallese, Miskito, Mixe (Coatlán), Mixtec (San Miguel el Grande), Mizo, Mooré, Mordvin (Erzya), Moru, Motuna, Murle, Navajo, Ngäbere, Ngambay, Nicobarese (Car), Nunggubuyu, Ojibwa (Eastern), Ossetic, Papiamentu, Piro, Pitjantjatjara, Pohnpeian, Polish, Purépecha, Quechua (Imbabura), Romani (Kalderash), Romanian, Romansch (Sutsilvan), Russian, Saami (Northern), Samoan, Sango, Santali, Seychelles Creole, Shilluk, Sora, Sougb, Spanish, Sranan, Swahili, Swedish, Tabassaran, Tagalog, Tajik, Tamil, Thai, Tibetan (Written), Timorese, Tlapanec, Toaripi, Tobelo, Tok Pisin, Tongan, Trique (Chicahuaxtla), Turkish, Udmurt, Ulawa (Sa'a), Uma, Veps, Vietnamese, Warlpiri, Wolof, Worora, Yoruba, Zapotec (Isthmus), Zoque (Copainalá), Zulu

Table 3 displays a small portion of the data from the database. The full sample is given in Table 2. Example (5) is from the French text and contains two contextually embedded situations (underlined) which have been chosen as analytical primitives in the database, see also examples (1) and (2) above.

(5) French (Indo-European, Romance) [Mark 1:29]

Ils quittèrent Ø la synagogue et allèrent aussitôt à la maison de Simon et d'André...

Table 3: Extract from the underlying database

Situations	English	French	Hait.Cr.	HmongNjua	Italian	Mapudungun	Russian	Tobelo	TokPisin	Wolof
1:5	un=to	a	N	?	a	_	k=D	?	long	ci
1:9	from	de	N	peg	da	mew	iz=G	oka	N	N
1:10a	out=of	de	nan	huv	da	mew	iz=G	ile	N	N
1:10b	up=on	sur	sou	sau	su=da	mew	na=A	uku	long	ci
1:11	from	de	nan	sau	da	mew	s=G	?	#	N
1:12	in=to	dans	nan	tom	in	mew	v=A	ika	long	ca
1:14	in=to	en	nan	peg	in	N	v=A	ika	long	ca
1:17	after	avec	N	N	OBJ	PRO	za=I	PRO=N	N	ci
1:18	OBJ	ACC	avek	N	OBJ	PRO	za=I	PRO=N	N	ci
1:20	after	avec	avek	N	dietro=a	PRO	za=I	PRO=N	N	ci
1:21a	in=to	a	nan	huv	a	N	v=A	ika	long	N
1:21b	in=to	dans	nan	huv	in	mew	v=A	ika	long	ci
1:25	out=of	de	sou	huv	da	OBJ	iz=G	de	N	ci
1:26	out=of	de	_	_	da	mew	iz=G	oka	N	ci
1:29a	out=of	N	N	huv	da	mew	iz=G	N	N	ci
1:29b	in=to	a	N	tom=tsev	in	mew	v=A	ika	long	ci

The database does not contain any diacritic signs. N: zero; _: Clause does not contain corresponding local referent phrase; #: Corresponding clause missing; ?: Unclear/not coded; PRO: head marking on verb; =: separates components.
Missing cells in the database (not attested, unclear) less than 8 %. Datapoints in total: 26'967 (all coded manually).

The distance matrix is computed by using Hamming distance as a distance measure.[6] For any pair of situations the number of differences in languages is divided by the total number of languages where both values are attested, which results in a distance matrix of 190 · 190 cells. To exemplify this only for the data given in Table 2, the situations 1:25 and 1:26 have a distance value of 2/8, because of the eight attested pairs two (Mapudungun, Tobelo) are different. For the pair 1:21b and 1:25 the value is 2/10=0.2 because only two texts use identical coding means (Hmong Njua, Wolof).

While exemplar-based databases imply less commitment to a priori definitions of semantic domains, the choice of analytical primitives always implies commitment in several respects which cannot be avoided. Pertinent issues are notably the following:

Sampling of analytic primitives: The 190 situations used here have been chosen from a larger set of 360 motion event clauses in Mark so that there are a large number of overtly

[6] Named after Richard Hamming who introduced it in the context of error-detecting and error-correcting codes (Hamming 1950).

expressed local phrases in order to avoid many not attested cells in the database. It is important to note that the dataset is biased toward certain domains as every typological dataset is. The semantic roles represented are goal ("to/into/onto"), source ("from/out of"), path ("along/through"), and companion ("following/going after/before"), while residence[7] (also called "locative" or place, place at rest, "in/on/at") is not represented (does not occur in motion events). The semantic roles are not represented with equal frequency, but rather with the frequency they happen to occur with in the particular text; thus in Mark goal is more frequent than source and source is more frequent than path. This raises the problem of sampling of situations. For some approaches it might be desirable to sample situations with less bias toward certain domains, but this is not possible without a commitment to semantic domains, such as, for instance, local roles. Moreover, when working with parallel texts, choice is restricted. Only situations which happen to be represented in the text can be chosen.

Delimitation of the set of forms considered: Given that adpositions grammaticalize gradually from nouns and verbs, there are no neat cut-off points even if we avoid the notoriously non-applicable distinction between adposition and case (see Kilby 1981). Here forms are excluded if they clearly derive from verbs (a motion verb with the same form still exists in the language).

Identity of forms: In many languages there are complex adpositions or local phrase markers consist of adposition and case, which both contribute to spatial semantics. Here complex local phrase markers are separated by equals signs (=), which allows the program that calculates the distance matrix to calculate several matrices making different choices. In the first map built in this section, partially identical forms are counted as halfway identical. Thus, for instance, Italian *a* and *a* are 100% identical, *a* and *dietro=a* are 50% identical and *a*

[7] The term "residence" may sound unusual, but I use it because it is more precise than "locative", "location" or "place" which are too ambiguous to denote the semantic role of a place at rest.

and *da* are 0% identical.[8] Section 4 considers how the map changes if decisions about identity are made differently and one major advantage of the program used here is that there is not one but three distance matrices calculated which can then be visualized as semantic maps.

To a certain extent, commitment is due to the fact that semantic maps are not built fully automatically. Ideally, a semantic map built from parallel texts would take whole translations of a text as input and build a semantic map of all token situations represented fully automatically. Automatic alignment has made much progress (see, e.g., Cysouw, Biemann and Ongyerth 2007) as far as wordforms are concerned; the problem is automatic morpheme analysis or algorithmic morphology (e.g., Goldsmith 2001) which has not reached a stage yet that it might be recommended for semantic map approaches. Moreover, dealing with fully automatic building of semantic maps would imply to have tools which can generate distance matrices and visualizations of several thousands of analytical primitives, which is a problem in itself.

From the database as illustrated in Table 3 the semantic map is built fully automatically. The distance matrix is calculated by a simple Python program which I programmed myself (Appendix). The matrix is then visualized by classical multidimensional scaling (the function cmdscale() in R, http://www.r-project.org). While there are many ready made tools for MDS from databases directly, there is reason for typologists to engage in programming the calculus of distance matrices, because this allows for generating several distance matrices from the same database with slightly different decisions about identity made (Section 4). The Python program has the further advantage that it generates a file with R-code which can be copied

[8] Distinguishing simple from complex forms is not strictly possible. Complex forms gradually merge in grammaticalization (for instance, French *dans* from Latin **de intus*) and there are numerous instances in the database where it can be discussed whether equals signs should be added or omitted (for instance, French *auprès*). While distinguishing simple from complex forms will therefore never be an ideal solution, it is argued in Section 4 that it is a more optimal solution than to disregard the distinction.

into the R Console to plot maps of the major categories in all doculects of the sample automatically.

The distance matrix is visualized by multidimensional scaling (MDS). MDS takes a matrix of pairwise dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities. If there are n items (analytic primitives), there is a maximum of n-1 dimensions (two dots can always be represented in one dimension, three dots can always be represented in two dimensions, etc.). The points are arranged such that the representation on the first dimension is as accurate as possible (as much information as possible is represented on the first dimension). Next the second dimension covers as much as possible of the information left and so on.

The dimensions are numbered but unlabeled and require interpretation. Before we consider the result of the MDS-analysis let us therefore compile a list of a priori semantic dimensions which might emerge in the analysis. There are a large number of possible semantically motivated formal distinctions in local phrase markers and it is assumed that at least some of them will emerge as dimensions in the MDS-analysis. Lower numbered dimensions are more relevant (account for more data in more doculects in the database). A difficulty is that many a priori "dimensions" are complex, i.e. do not lend themselves to a geometrical representation on a single dimension. If we take the example of local roles, there are at least five major subdomains: source (from, out of), goal (to, into, onto), residence (at, in, on), path (along, through), and companion (with, following after, preceding before). However, matters are simplified by the fact that not all local roles are represented in the database. Since the data is restricted to motion events, the role residence is not represented. Table 4 gives a non-exhaustive list of potential semantic dimensions in local phrase markers.

Table 4: Expectable ("a priori") semantic dimensions in local phrase markers

Role	Source, goal, residence, path, companion (e.g., Fillmore 1971/75:26, Wälchli and Zúñiga 2006, Kibrik 1970, Ganenkov 2002)
Animacy	To a place vs. to a person (further distinctions for 1st, 2nd vs. 3rd person and/or proper names), honorifics (Korean)
Localization/Topology	Interior/containment (empty/full), top/support, proximity, contiguity (in, on, at, under, etc.; Levinson and Meira 2003, Wälchli and Zúñiga 2006, Kibrik 1970)
Absolute frame of reference	Northward, southward, seaward, landward etc.
Relative frame of reference	In front, behind, etc.
Transitivity	Object/absolutive vs. oblique
Definiteness	to the house vs. to a house, etc.
Deixis	Ground here vs. ground there, etc.
"Altitudunal cases"	Low, Level, High (Rai languages, not represented in the sample; Ebert 1999)
Classification on the basis of ground	E.g., into liquid vs. into fire, etc.
Proper name	Place name vs. appellative
Distance	Close distance vs. extreme distance (Hopi: Malotki 1979)
Generality	Omnipurpose oblique markers vs. specific markers (Comrie 1986)
(Demonstrative) Adverbs behaving differently	"Thence", "thither", "hence", "home"
Lexicalization with particular verbs	E.g., "enter" with residence or goal

The dimensions listed in Table 4 are not restricted to spatial semantics in a narrow sense. Any recurrent formal difference in local phrase markers can be relevant. Thus, demonstrative adverbs often have a different form from local phrase markers on nouns and certain verbs, such as "enter" can require particular local phrase markers.

Let us now consider the constellation of places how it emerges in the MDS analysis and how it is instantiated in a number of doculects of the database. The languages in the discussion below are chosen such that many different category types are covered in order to illustrate the range of diversity attested. A summary of the results for the a priori dimensions listed in Table 4 i given at the end of this section.

It turns out that for this particular dataset only the three first dimensions correspond to interpretable semantic distinctions. In this respect, local phrase markers differ from lexical verbs where many more dimensions can be interpreted (Wälchli and Cysouw forthc.). Figure

3 plots the first two dimensions and illustrates the semantic map with French categories (top-left). The dots are the 190 analytic primitives as arranged by the MDS analysis in Dimension 1 (x-axis) and Dimension 2 (y-axis). The symbols are assigned according to the local phrase markers present in the parallel text depicted. The Python program in Appendix A writes a code for the program R which produces these plots for all parallel texts automatically from the database. The categories are arranged according to their frequency. Thus, *dans* happens to be the most frequent local phrase marker in the French text in the 190 situations considered, followed by *de* and *a*. The number of categories maximally represented is limited to eleven and to categories occurring at least twice. The small grey circles are situations which are not represented by any category matching these criteria for the doculect plotted (rare categories or situations which happen to be not attested in the database for the particular parallel text).

The MDS Dimension 1 can be interpreted as corresponding to the a priori dimension role. It distinguishes very neatly source (negative values[9]) and goal (positive values) with path being intermediate. The absence of the role residence illustrates the importance of the choice of analytic primitives. The map would change if situations representing residence were added.

The reason why the source-goal distinction clearly emerges is that there are many doculects in the sample like French where the major categories are more or less strictly sensitive to the source-goal distinction. In French there are a few outliers for the source preposition *de* on the goal side due (a) to the verb *s'approcher de* 'approach' and (b) the expression *de l'autre côté* 'to/at the other side'. In the database it is rare that these two particular subdomains are marked the same way as source which is why the few situations having *de* with goal are outliers. This reflects the fact that it is unlikely (but not impossible) that a language picked at random will combine source and "approaching" and "other side" in

[9] The orientation of the poles is completely accidental in MDS.

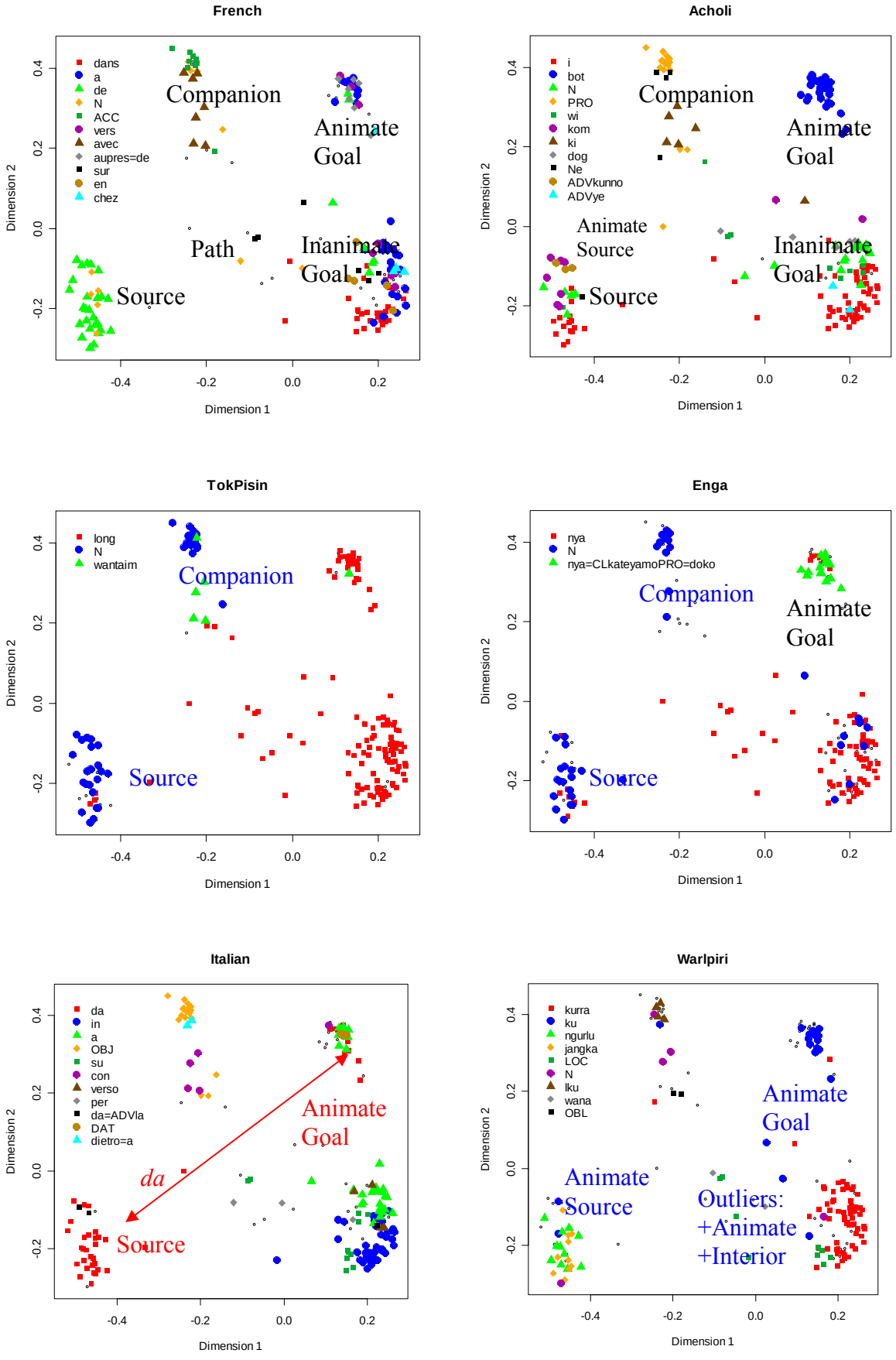
a single category. However, it is very likely that a language picked at random combines all or various source situations in the same category. A traditional semantic map would abstract away from minor anomalies such as *de l'autre côté*. Including the effect of all such minor subdomains renders it impossible to draw universal or implicational maps. Universal semantic relationships emerge only at a high level of data reduction. On the level of exemplars there are hardly ever strictly scalar relationships. Probabilistic semantic maps have the advantage that they do not require any previous idealization of the data. The general trends in the data clearly emerge even if there are many minor outliers.

A major aim in typology is to identify strong general tendencies within the whole range of cross-linguistic diversity. Many methods of typology do this at the cost of heavy data reduction as is expressed by the very name "typology": languages and semantic domains are forced into given sets of types. Probabilistic semantic maps are a more empirical and less idealizing typological tool. In probabilistic semantic maps we can identify major trends in the data without abstracting away from more idiosyncratic aspects. It is a method for doing "typology without types". Figure 3 shows that French contributes to the general trend of source-goal distinction in local phrase markers, but not without exceptions.

Dimension 2 represents not only a single a priori dimension, but a combination of two. The MDS analysis arranges the situations such that as much information in the database supports the first dimension, next from the information left as much information as possible is represented in the second dimension and so on. If a priori dimensions are "orthogonal"—which means that there is no or little interaction between them—there is no way to combine them in one dimension. However, if two a priori dimensions can be combined in the same probabilistic "scale" or pseudo-scale, MDS analyses will tend to combine them and this is what happens here. Localization has many more than two poles, but for motion three are dominant: proximity, surface, and interior. These form a kind of contact scale: interior is a

closer contact than surface and proximity is lack of contact. Animate goals usually occur with localization proximity. Motion into or onto a person is more rarely expressed than motion to a person in motion events. This makes it possible to add animacy at the loose contact end of the contact pseudo-scale. It has to be emphasized, however, that this is no absolute but only a probabilistic scale which is supported by a large amount of data in the database, but not without exceptions.

Figure 3: Semantic map of local phrase markers in French, Acholi, Tok Pisin, Enga, Italian, and Warlpiri



(6) Emergent probabilistic scale combining animacy and localization in Dimension 2 in Goal contexts:

animate (proximity)	> inanimate proximity	> neutral localization (to place name)	> inanimate surface	> inanimate interior

This pseudo-scale is illustrated nicely by Figure 3 (top-right) for Acholi (Crazzolara 1955): *bòót* 'to, from side (animate)' > *köòm/koòm* 'body, on' > *dóg* 'mouth, bank, to' > Zero (mostly with place names) > *wiìc* [*wiì-*] 'head, top, on' > *ì* 'inside' (from *iìc* 'belly'). Figure 3 (top-right) also shows that Acholi does not at all distinguish role in local phrase markers. It also shows that animate source happens to be very weakly represented in the database. The few examples with animate source happen to be expressed with *köòm/koòm* 'body, on' and PRO (transitive verb with head marking, one situation only).

The third dimension, which is not plotted in the figures, distingushishes animate goal from companion in the two poles with all inanimate ground situations being intermediate. Companion ("following after somebody, preceding somebody, go with") and animate goal are both on the animate pole of Dimension 2. However, they are distinguished already by Dimension 1 where animate goal goes together with inanimate goal and companion exhibits a slight affinity to source, which is due to languages such as Tok Pisin (Figure 3 middle-left), where both companion and source tend to be expressed by transitive verbs (*bihainim* 'follow', *lusim* 'leave, exit'), which is why they share the category zero marking (N). While this combination of source and companion by means of transitivity and zero marked ground phrase is dominant in Tok Pisin and other languages of New Guinea, such as Enga (Figure 3 middle-right) it occurs to a lesser extent also in some European languages, such as French (object of *quitter*, *suivre*; accusative with pronouns, zero ["N"] with nouns). In Enga, animate

goal is expressed by a subordinate clause with the verb *katenge* 'be' ("where somebody is"), illustrated in (7).

(7) Enga (Trans-New Guinea, Engan) [Mark 10:13]

Wane	wanaku-pi	namba	**ka**-ly-o	doko-nya	epena
boy	girl-PL	I	**be**-PRS-1SG	that-LOC	come-3SG.IMP.IMMED
daa	lao	kaita	lyok-ala	naeya-lapa-pe.	
not	want	path	break-PURPOSE	take-IMP.LATE-2PL	

'Suffer the little children come unto me...'

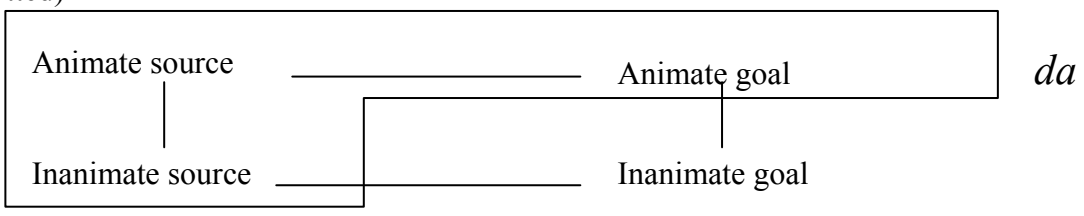
In Italian (Figure 3, bottom-left) *da* (source) is distinguished in Dimension 1. Dimension 2 makes a rather neat localization distinction between *in* (interior) and *a* (non-interior) within the goal cluster. However, *da* source has a long-distance connection to animate goal (...*essi andarono da lui* '...they came unto him' Mark 3:13). The long distance on the probabilistic semantic map reflects the fact that the categorization pattern of Italian *da* is rare (a parallel is Gbeya *ha*, Samarin 1966:73). The closer the dots in a language-particular category cluster, the more this category is recurrent cross-linguistically.

Italian *da* can be better represented on a traditional map (Figure 4), which focuses on the particular semantic similarity relationships relevant for Italian *da* and abstracts away from all potential latent similarity relationships, which would also require a different set of analytical primitives. For instance, Figure 4 abstracts away from the localization difference between interior and non-interior which is relevant for the distinction of *a* and *in* but is irrelevant for

da. This example illustrates the difference between probabilistic and traditional maps and shows that the two types of semantic maps have complementary functions.

From the point of view of probabilistic maps, implicative maps represent semantic space where all conflicting evidence is removed from the focus of attention, such that general tendencies emerge in a clear and pure form. In the probabilistic map based on usage data, animate source does not emerge as a cluster since animate source is less often distinguished in local phrase markers from inanimate source than animate goal from inanimate goal and since animate source happens to be a rare context in the particular text Mark considered. Traditional maps are not sensitive to such distortions by usage and can reflect language-particular systems more accurately, they are more schematic. Probabilistic maps are completely indifferent to language-particular systems and are completely usage-based. No semantic map can reflect all potential similarity relationships in such a domain as local phrase markers. Probabilistic maps privilege frequent categorization patterns, rare categorization patterns are better represented on language-particular semantic maps which abstract away from all latent semantic distinctions which are irrelevant for a particular category.

Figure 4: Implicational map of Italian *da* (only source and goal given, the role residence is omitted)



The Italian example also illustrates that the probabilistic map does not necessarily represent accurately all semantic distinctions which are associated with the MDS dimensions. Even though Dimension 1 reflects role and Dimension 2 reflects animacy, this does not imply

that all aspects of role and animacy would be well represented, not if they are rare, such as the Italian *da* source-and-animate-goal connection.

Warlpiri and Pitjantjatjara illustrate similar points. In both Waripiri and Pitjantjatjara there are aspects of usage concerning the animacy distinction, which are not conforming the general trends in the dataset, even though they are well in line with the animacy hierarchy in abstract terms. Warlpiri (Figure 3, bottom-right) makes a distinction between source and goal only for inanimate, not for animate goals, where the Dative (*-ku/ki*) is used. The animate category is stricter in Warlpiri than in most other languages. The contexts of going into and coming out of an animal or person go together with animate (that is, Dative, example 8), while in most languages with an animacy distinction these contexts go together with inanimate. This is reflected on the probabilistic semantic map for Warlpiri by some outlier exemplars distant from the animate cluster for the Dative *-ku/ki* category.

(8) Warlpiri (Australian, Pama-Nyungan) [Mark 5:13]

*...**wilypi=pardi**-ja* *wati-**ki**,* *yaarl=yuka-ja-lku-lu-jana*

PV=exit-PST man-**DAT** PV=enter-PST-then-3PL.SUBJ-3PL.OBJ

*nguurrnguurrpa-**ku**-ju.*

pig.PL-**DAT**-EMP

'[And the unclean spirits] went out, and entered into the swine...'

A further distinction in Warlpiri *-jangka* elative of origin vs. *-ngurlu* elative is uncommon in the sample.

Pitjantjatjara (Goddard 1996, Figure 5, top-left) has special forms for pronouns and names including place names with an element -*la/ta/ṯa/tja* for all local cases (allative -*kutu* vs. -*lakutu*, locative -*ngka* vs. -*la*, ablative -*nguṟu* vs. -*languṟu*, perlative -*wanu* vs. -*lawanu*). Proper names are often higher on the animacy hierarchy than appellatives, but prioritizing names including place names over animacy proper results in a rare categorization pattern in usage which is reflected by discontinuous representations of the categories on the semantic map. The high number of Locative in Pitjantjatjara is due to the construction of the enter/arrive verb *tjarpanyi* with locative rather than allative.

Another rare category connected to the animacy scale is the honorific animate goal marker in Korean (honorific animate goal -*kkey* vs. Animate goal -*eykey*; Chang 1984:196). The honorific -*kkey* happens to be frequently represented in Mark because Jesus (honorific) is a recurrent animate goal (Figure 5, top-right). No other language of the sample makes a similar distinction in local phrase markers.

Let us now consider some examples of rare categorization patterns beyond the dimensions of role, animacy and localization in Tobelo, Wolof, and Hopi.

According to Holton (2003:34-35), Tobelo has (a) a Locative suffix -*oka*, (b) Allative (-*ika*, "motion toward the noun") and Ablative (-*ino*, "motion away from the noun") suffixes, (c) a first dimension of directional suffixes seaward (-*óko*) vs. landward (-*iha*), (d) a second dimension of directional suffixes -*úku* 'down' vs. -*ilye* 'up', and (e) zero marking ("N", directional suffixes are not obligatory). Furthermore, there is a preposition *de* 'with, and' (Holton 2003:30), used in the N.T. doculect also for some cases of animate source. In the N.T. Tobelo doculect (Figure 5, middle-left) the Locative -*oka* is used in some goal and

especially some source contexts, the Allative *-ika* is restricted to goal contexts. This can be interpreted such that the Locative is more general than the Allative. However, the Ablative *-ino* does not behave as expected for a source marker; *-ino* is attested for source and goal contexts. As Holton (2003:47) points out, *-ino* is also a directional suffix of verbs 'toward (ALLATIVE)'. The examples in Mark document that *-ino* is sensitive to deixis even in adpositional use. In (9a) with a first person ground *-ino* marks a goal, in (9b) with a third person ground the source is marked by Allative *-ika* (for second person in Mark 9:19 there is *-ika* as well).

(9) Tobelo (West Papuan, North Halmaheran)

a. Ni-ao o ngohaka gënanga neng-**ino**

2PL-bring NM child that PROXIMAL.PUNCTUAL-ABL

'...bring him unto me.' [Mark 9:19]

b. ... iwi ao o ngohaka gënanga o Yesus-**ika**

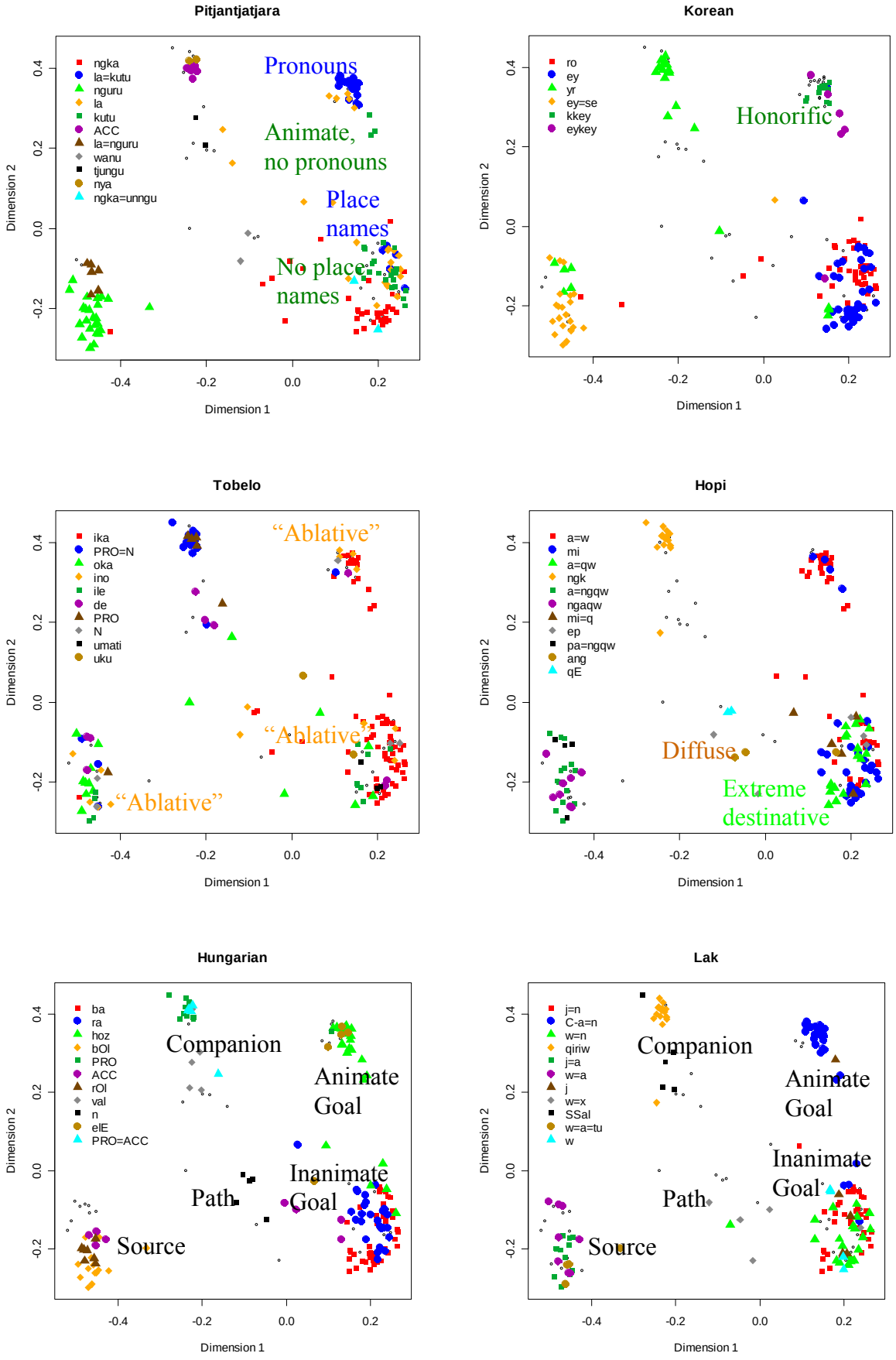
3PL>3SG.M bring NM child that NM Jesus-ALL

'And they brought him unto him.' [Mark 9:20]

"Ablative" for goal is not restricted to first person; there are also some examples for place where the first person is or for a deictically closer third person. While the consideration of the Tobelo Mark examples is not sufficient to describe the exact usage of Ablative *-ino* (and there would be more to say about the category; for example, *-ino* also marks path in several

examples), what we can clearly see from the map is that Allative *-ika* is a canonical goal marker from a typological point of view while the "Ablative" *-ino* is not a canonical source marker, but a rare category which is not supported by the rough semantic grid provided by Dimensions 1 and 2. This example illustrates that in a pair of markers within a language-particular system one of the markers can be semantically common and the other one exotic from a cross-linguistic point of view.

Figure 5: Semantic map of local phrase markers in Pitjantjatjara, Korean, Tobelo, Hopi, Hungarian, and Lak



Tobelo is not the only language in the sample with a local phrase marker with some deictic semantic component. However, there is no other category in any language with a closely similar range of use. Another example for deictic local phrase markers is the pair of Wolof omnipurpose oblique prepositions *ci* 'proximal' and *ca* 'distal' (see, e.g., Robert 2006). The map for Wolof (Figure 1, right) shows that Dimensions 1 and 2 are not sensitive to the semantic distinction between these two prepositions. However, an interesting minor generalization in the N.T. doculect is that *ci* proximal goes together with '(enter) into'-contexts (bottom of goal inanimate cluster), whereas *ca* distal is preferred in this narrative text with '(go) to'-contexts. Entering usually implicates a shorter (proximal) journey and going to a longer (distal) journey, as exemplified in examples (10a/b):[10]

(10) Wolof (Niger-Congo, Northern Atlantic)

a. ...ñu njël-u dem **ca** bàmmeel ba

3PL dawn-MIDDLE go **PP:DIST** grave DEF:DIST

'[And very early in the morning the first day of the week,] they came unto the sepulchre [at the rising of the sun.]' [Mark 16:2]

b. Ñu dugg **ci** bàmmeel bi...

3PL enter **PP:PROX** grave DEF:PROX

'And entering into the sepulchre...' [Mark 16:5]

[10] Other Wolof markers in Figure 1 (right) are Zero "N", OBLique form (pronouns only), *ak* 'and, with', *fi* 'here' (partly used in a preposition-like way), *fa* 'there', *ci kaw* 'on top (proximal)'.

However, there is another language in the sample, Hopi, with a kind of distal-proximal distinction, termed Extreme (vs. Non-extreme), where entering behaves exactly the other way round: interior contexts in Hopi are extreme, surface and proximity contexts tend to be Non-extreme (if there is no other reason to mark them as extreme such as in 11, where top is an unusual location). Examples (12a/b) illustrate the subtle contrast between the top and the interior of a fireplace, distinguished by Extreme vs. Non-extreme.

(11) Hopi (Uto-Aztekan, Hopi; Malotki 1979:92)

Nu? kits?o-ve-q-ni-q su-?inu-mi-q tatsi tso?ó-m-ti

I roof-PUNC-EXTR-NEX-DS right-I-DEST-EXTR ball jump-MULTI-RE

'I was on the roof, and the ball jumped up to me'

(12) Hopi (Uto-Aztekan, Hopi)

a. Nu? ?a-w ?öngáp-ta

I there-DEST cooked.beans-CAUS

'I put on beans (for cooking)' (Malotki 1979:35)

b. ?a-qw qöö`na-?a

there-DEST.EXTR firewood-IMP

'Put more firewood into the fire!' (Malotki 1979:62)

Hopi (Figure 5, middle-right) has a complex system of local phrase markers discussed in great detail in Malotki (1979) and summarized in Table 5. The major language-particular semantic dimensions at work are (a) role (goal: Destinative, source: Ablative, place/path:

Punctual and Diffuse), (b) Extreme vs. Non-Extreme, (c) Punctual vs. Diffuse, (d) Case vs. Postposition with a "reference basis". In addition there are proximal and distal adverbs.

Table 5: System of Hopi local cases and adpositions according to Malotki (1979), simplified

	Case	"Reference basis" ?a-+ Postp.	Proximal ("here")/distal("there")
Punctual	*-pe/ve*	*?e-p*	*ye-p/pe-p*
Extreme punctual	*-pe-q/-ve-q*	*?e-pe-q*	*ye-pe-q/pe-pe-q*
Diffuse	*-pa/-va*	*?a-ng*	*ya-ng/pa-ng*
Extreme diffuse	*-pa-qe/-va-qe*	*?ang-qe*	*yang-qe/pang-qe*
Destinative	*-mi*	*?a-w*	*yuk/panso*
Extreme destinative	*-mi-q*	*?a-qw*	*yukyiq/panso-q*
Ablative	*-ngaqw*	*?a-ngqw*	*ya-ngqw/pa-ngqw*

Whereas the role dimension is cross-linguistically common and therefore neatly mapped on Dimension 1, the proximal and distal adverbs play a minor role. However, the Extreme vs. Non-extreme and the Punctual vs. Diffuse distinctions are cross-linguistically rare, maybe unique in their concrete manifestation. Accordingly, they do not emerge as clusters in the probabilistic semantic map (it is unlikely to encounter a language like Hopi if one language is picked at random).

The Punctual vs. Diffuse distinction is illustrated in (13a/b) and is often connected with presence (13a) or absence (10b) of a distributive component.

(13) Hopi (Uto-Aztekan, Hopi)

a. Nu? ?a-ng soòso-k saavu-t poò-pongi

I there-DIFF all-ACC wood-ACC red-pick.up

'I have picked up all the (hackled) wood.' (Malotki 1979:52)

b. ?uù-?aya-y ?e-p kwusu-?u

POSS2SG-rattle there-PUNC pick.up-IMP

'Pick up your rattle!' (Malotki 1979:52)

However, Diffuse is also generally used for path: "Jede Linienvorstellung, sei sie statisch-konkret als visuelles Phänomen gegeben oder dynamisch-abstrakt als linearer Bewegungsablauf, wird im Hopi diffus gedeutet [...] Die Vorstellung 'entlang' resultiert in typischer Weise aus einer Linieninterpretation, die an einem langgestreckten Bezugsort vorbeiführt" (Malotki 1979:55).[11]

One might be inclined to believe that 190 situations from a narrative text would be enough to represent the range of functions that can be expressed by local phrase markers in motion events. However, given the large number of possible distinctions, this is not the case; especially because many situations express very similar situations (the situations such as they occur in a narrative text are not semantically equidistant). If we consider languages with moderately large or large case systems, such as Hungarian and Lak, not all cases are represented. In Tables 6 and 7 the cases occurring in the 190 situations are marked boldface. The semantic maps of Hungarian and Lak are given in Figure 5 (bottom).

Table 6: Hungarian local cases

	SOURCE	RESIDENCE	GOAL
IN	***-ból/ből***	***-ban/ben***	***-ba/be***
ON	***-ról/ről***	***-n/on/en/ön***	***-ra/re***
AT	*-tól/től*	*-nál/nél*	***-hoz/hez/höz***

[11] "Every concept of a line, be it given as a static-concrete visual phenomenon or as a linear movement is interpreted as diffuse in Hopi [...] The concept 'along' results typically from the construal of a line parallel to an extended ground." (translation BW).

Table 7: Lak case system (Xajdakov and Žirkov 1962)

		Local case series					
Nom.	*kkatta*	I	Loc.	*kkatluwu* ‘in’	IV	Loc.	*kkatlulu* ‘under’
Gen.-Erg.	*kkatlul*		Lat.	***kkatluwun***		Lat.	*kkatlulun*
Dat.	*kkatlun*		All.	*kkatluwunmaj*		All.	*kkatlulunmaj*
Abl.	*kkatluša*		Prosec.	***kkatluwuχ***		Prosec.	*kkatluluχ*
Comit.	*kkatlušal*		Abl.	***kkatluwa(tu)***		Abl.	*kkatlula(tu)*
Comp.	*kkatlujar*	II	Loc.	*kkatluj* ‘on’	V	Loc.	*kkatluč’a* ‘near’
„because“	*kkatluχlu*		Lat.	***kkatlujn***		Lat.	***kkatluč’an***
Sociat.	*kkatlujnu*		All.	*kkatlujnmaj*		All.	*kkatluč’anmaj*
‘at’	*kkatlux*		Prosec.	***kkatlujχ***		Proseq.	*kkatluč’aχ*
‘to’	*kkatluxxun*		Abl.	***kkatluja(tu)***		Abl.	*kkatluč’a(tu)*
		III	Loc.	*kkatluχ* ‘behind’	VI	Loc.	*kkatluc’* ‘at very’
			Lat.	*kkatluχun*		Lat.	*kkatluc’un*
			All.	*kkatluχunmaj*		All.	*kkatluc’unmaj*
			Prosec.	*kkatluχuχ*		Prosec.	***kkatluc’uχ***
			Abl.	*kkatluχa(tu)*		Abl.	*kkatluc’a(tu)*

After having considered how languages with different systems of local phrase markers are represented in the semantic map built here we can conclude that the following of the a priori semantic dimensions listed in Table 4 are represented and hence represent general trends in local phrase markers cross linguistically. Role is represented in Dimension 1 (source – path – goal), but also partly in Dimension 2 (companion). Animacy is represented in Dimension 2. However, animacy is not equally well distinguished for all roles; it is distinguished especially within the role goal and less clearly in source since animate source is less frequently represented in the database and less frequently distinguished from inanimate cross-linguistically. Topology (interior, surface, proximity) is represented to a certain extent in Dimension 2 in combination with animacy in a probabilistic “degree of contact” scale. Interior contexts are slightly more central on Dimension 1 (role), because many languages construct ‘enter’ verbs with residence rather than with source markers. Transitivity plays a minor role for arranging companion and source closer to each other than companion and goal.

Generality is not represented as dimension but as the spread of a category over a larger area of the map. However, categories spread over larger areas of the map can also be cross-linguistically rare categories not supported by any other language of the sample. Most other dimensions listed in Table 4 do not emerge as dimensions or clusters on particular dimensions. Put differently, the semantic map built here is no good tool to appropriately represent the categorization systems in all languages. However, it is a good tool to compare a large number of languages directly on the level of language use and to distinguish general recurrent trends from more specific language-particular categories.

4 Variations without a theme: how different samples and different ways to count can change a semantic map

According to Haspelmath (2003:217) "[e]xperience shows that it is generally sufficient to look at a dozen genealogically diverse languages to arrive at a stable map that does not undergo significant changes as more languages are considered." This claim can easily be shown to be wrong for probabilistic semantic maps. Let us take a subsample from the 153 languages containing 42 languages from 18 families (according to the WALS classification) and some creole languages (Table 8).

Table 8: 42-language subsample:

Acholi, Adyghe, Akan, Ambulas, Bambara, Bari, Bribri, Cakchiquel, Choctaw, Creek, Efik, Ewe, Haitian Creole, Hmong Njua, Igbo, Ijo (Nembe), Ju|'hoan, Kabba-Laka, Kabiyé, Koyra Chiini, Kuna, Lahu, Mandarin, Mapudungun, Mixe (Coatlán), Mixtec (San Miguel el Grande), Mooré, Murle, Ngäbere, Ngambay, Nicobarese (Car), Ojibwa (Eastern), Purépecha, Sango, Seychelles Creole, Sranan, Swahili, Tlapanec, Toaripi, Trique (Chicahuaxtla), Wolof, Zulu

Figure 6 shows how the French and Acholi categories are arranged in Dimensions 1 of 2 of a MDS analysis based on the 42 doculects in Table 8. Whereas in Section 3 we have always

represented texts which have contributed to build the semantic map, here a doculect which has not contributed to the configuration of situations is shown. Put differently, we have constructed a model based on 42 languages and now consider whether this model is accurate to visualize also categories of other languages. The answer is no for French. The languages in Table 8 happen to be all like Acholi in that they do not encode the source-goal distinction in local phrase markers, which is why no role distinction emerges in the MDS analysis.[12] What we get is now animacy in Dimension 1 and animate goal vs. companion in Dimension 2 (further dimensions do not support any interpretable semantic distinctions). The zero marked class in Acholi, going together with companion in Dimension 2 raises an important problem of semantic maps. It happens to be the case that many languages lacking the source-goal distinction in local phrase markers have unmarked place names and it happens to be the case that the ground in "follow" companion contexts is often an object which in turn is often unmarked. The recurrent formal identity shared between companion and place names consists thus mainly of a lack of any marking. In the present approach, shared zero marking is counted the same way as any overt shared marker, even though zero marking is a much less characteristic formal property, so that it is highly doubtful whether shared zero marking is an argument for similarity in meaning (see Wälchli 2005:30 for discussion).

Figure 6 shows that if we happen to pick the "wrong" forty-odd languages from one and a half dozen language families, it can happen that we miss the most dominant world-wide trend in the data. This does not mean anything else than that sampling is a highly relevant issue for semantic maps, which is not much of a surprise, given that it is well known in typology and

[12] However, most of these languages distinguish source and goal in verbs, which are disregarded here. In Ewe it could be argued that the verbs encoding source and goal have grammaticalized to prepositions, but these elements are not coded as local phrase markers in the underlying database.

Interestingly, animate source is intermediate between inanimate and animate goal on Dimension 1, because animate source is more often not distinguished from inanimate than animate goal.

especially areal typology that sampling matters (see, e.g., Nichols 1992, Dryer 1989). Semantic maps are no exception. Building semantic maps is as sensitive to sampling as is any other typological method. Every sample of languages or doculects reflects a certain amount of cross-linguistic diversity which can serve as a basis to construct a model that applies to all language data which fall into the range of the structural diversity represented in the data underlying that model. This reminds us of the fact that the 153 language sample is a convenience sample with a strong bias toward European and Indo-European languages even if it contains languages from all continents.

Let us therefore build a model based on a more balanced subsample. The 84 doculects used are given in Table 9. Figure 7 (left) shows the French categories plotted on this map and Figure 7 (right) shows the differences between the 153 language and the 92 language sample maps in location between the situations plotted as lines.

Figure 6: Semantic map of local phrase markers based on languages without source-goal distinction

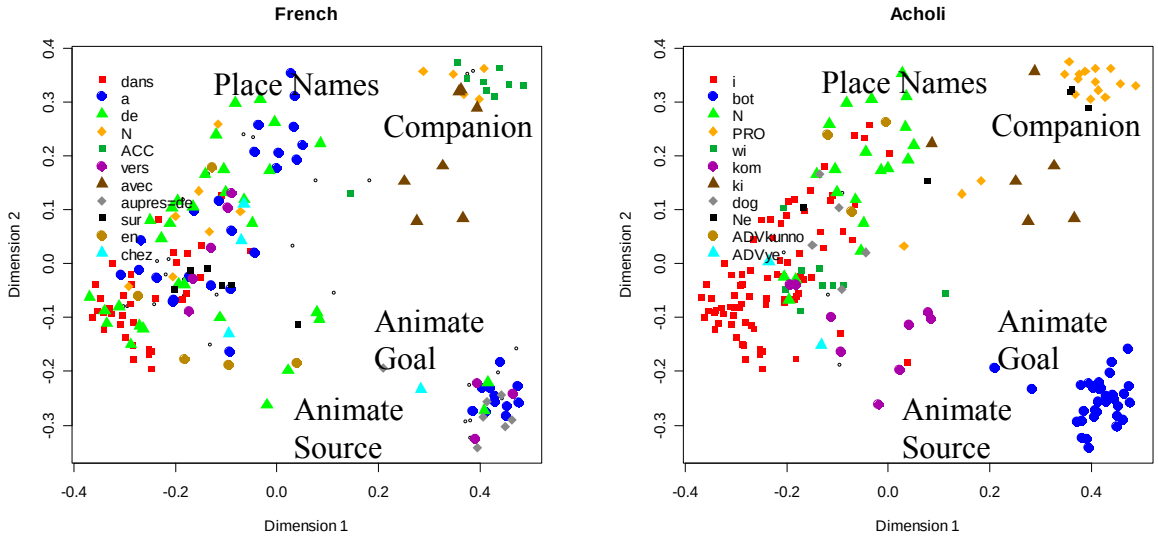


Table 9: 84-language subsample with reduced bias[a)]

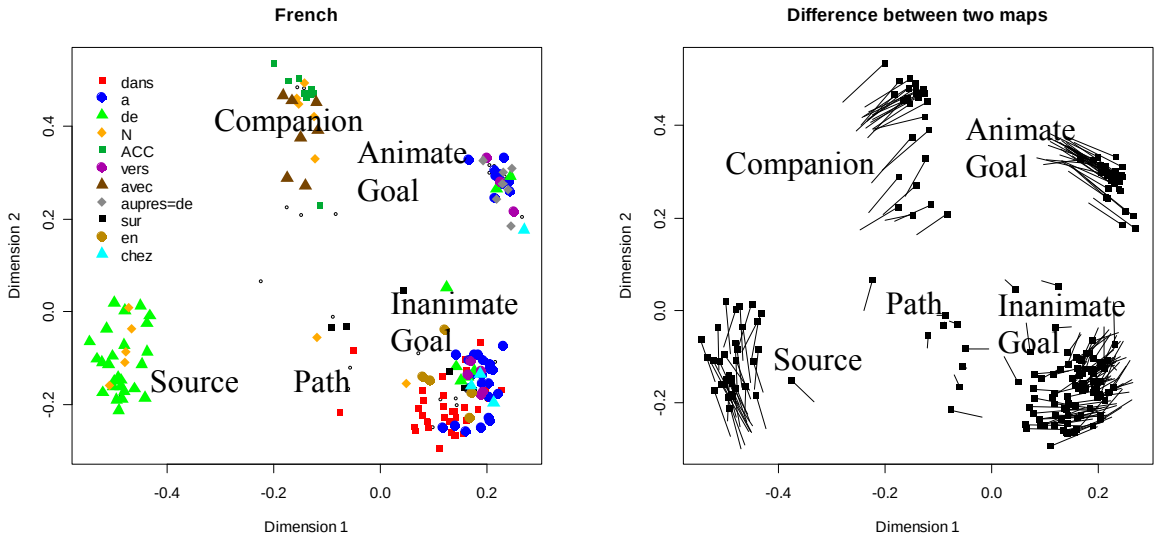
Africa [16]: Bari, Ewe, Gbeya Bossangoa, Hausa, Ijo (Nembe), Ju|'hoan, Kabba-Laka, Kabyle, Khoekhoe, Koyra Chiini, Kunama, Maltese, Moru, Murle, Swahili, Wolof;
Creole [2]: Papiamentu, Tok Pisin;
Eurasia [15]: Adyghe, Ainu, Avar, Basque, Breton, Georgian, Greek (Classical), Hindi, Kannada, Khalkha, Korean, Lak, Lezgian, Liv, Mari (Meadow);
SEA & Oceania [13]: Garo, Hmong Njua, Jabêm, Khasi, Lahu, Mandarin, Maori, Mizo, Nicobarese (Car), Santali, Thai, Timorese, Vietnamese;
New Guinea & Australia [15]: Ambulas, Enga, Kala Lagaw Ya, Kâte, Kiwai, Kuku-Yalanji, Kuot, Motuna, Nunggubuyu, Pitjantjatjara, Sougb, Toaripi, Tobelo, Warlpiri, Worora;
North & Mesoamerica [12]: Cakchiquel, Choctaw, Cree (Plains), Dakota, Hopi, Mixe (Coatlán), Mixtec (San Miguel el Grande), Navajo, Purépecha, Tlapanec, Zapotec (Isthmus), Zoque (Copainalá);
South America [11]: Amuesha, Aymara, Bribri, Chiquito, Guaraní, Kuna, Mapudungun, Miskito, Ngäbere, Piro, Quechua (Imbabura)

[a)] Languages are assigned to continents according to their membership to language families not to geographical continents, as in the case of Maltese which is African, because Afro-Asiatic is an African rather than European language family.

As can be seen from Figure 7, the difference between the two maps does not in this case alter the maps substantially, the dimensions remain the same.

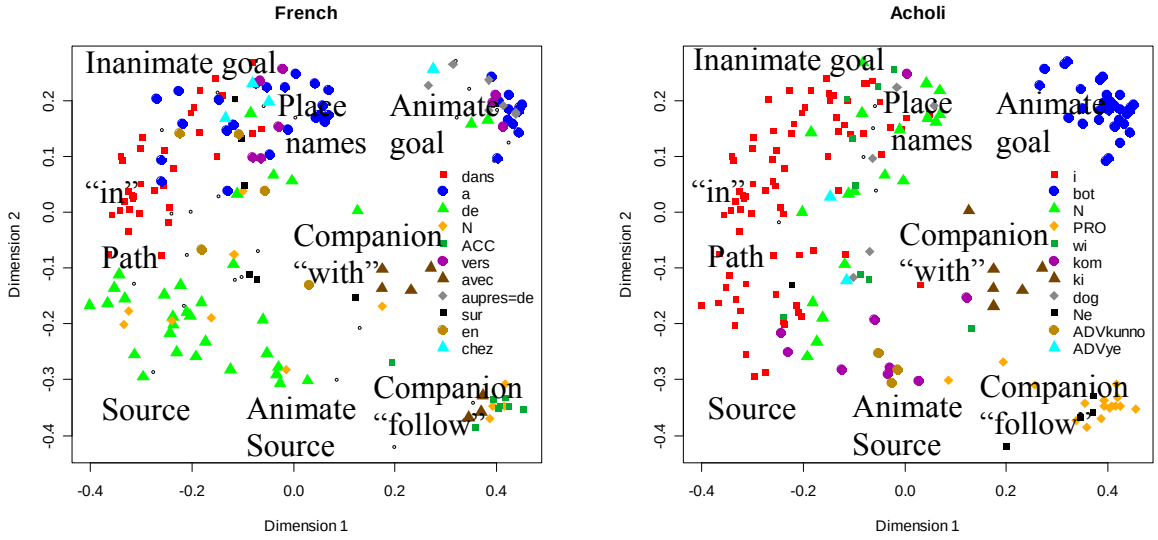
Figure 8 gives the semantic map for French and Acholi built on the basis of 27 African languages. The source-goal distinction emerges only in Dimension 2 and the distinction is not very marked. Dimension 1 is animacy/contact. Figure 8 shows nicely that the categories of an African language such as Acholi are better represented on a semantic map based on African languages than the categories of French.

Figure 7: Semantic map of local phrase markers with less biased sample in French and difference between the maps based on the more balanced 84-language sample (squared ends of lines) and the 153 language convenience sample (unmarked ends of lines)[a)]



[a)] The orientation of the y-axis has been inverted for the map based on the 84-language sample for better comparability

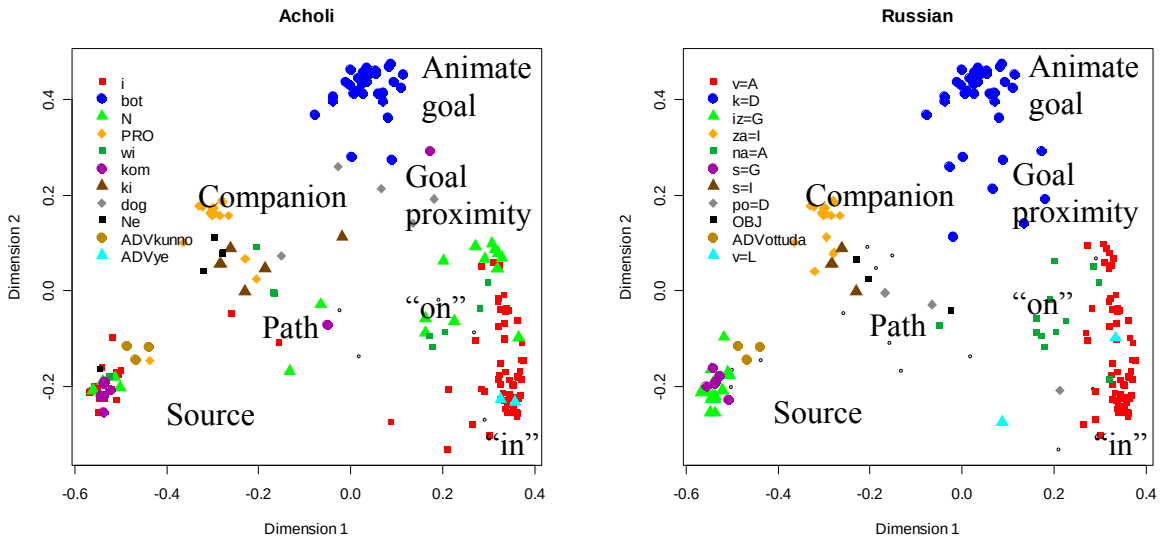
Figure 8: Semantic map of local phrase markers based on 27 African languages in French and Acholi



In the same way as mapping continent samples we can map family samples. If only the Indo-European languages are taken as a basis, the distance between inanimate and animate goal shrinks because there happen to be few Indo-European languages with an animacy distinction (Figure 9). This is illustrated with Acholi which makes a clear animate goal

distinction. Dimension 2 remains a probabilistic "degree of contact" scale when built on the basis of the twenty-seven Indo-European languages, but localization is more dominant now than animacy. The cluster for proximity, such as represented for instance by the Russian preposition *k* with Dative, becomes more compact on this Indo-European based map.

Figure 9: Semantic map of local phrase markers based on 27 Indo-European languages in Acholi and Russian

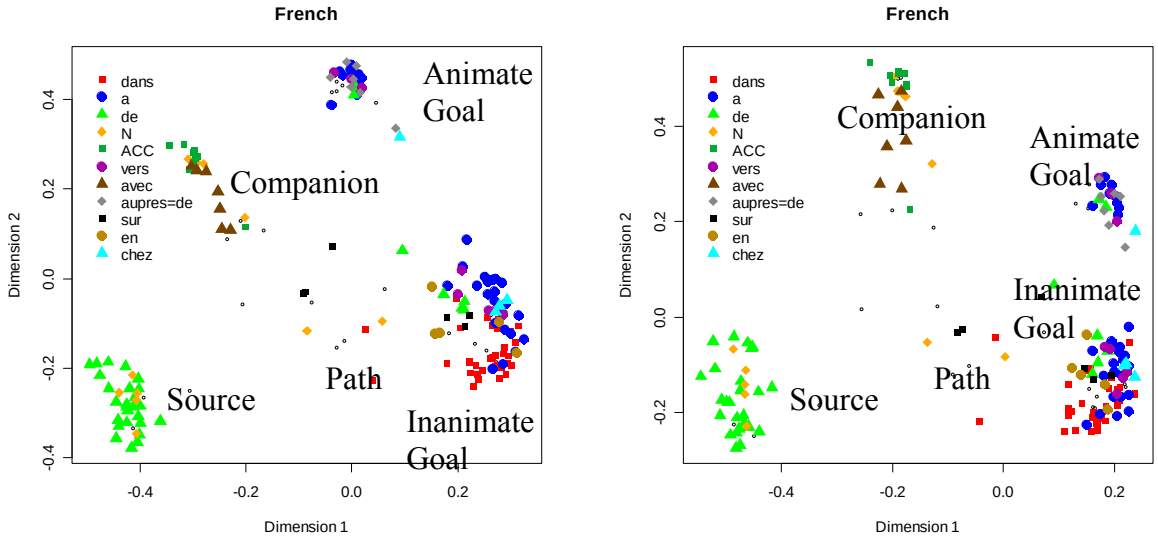


The program given in Appendix A calculates three distance matrices for an input data matrix. The three distance matrices are all calculated by means of the same distance measure Hamming, as discussed above. However, they differ as to how partially identical categories are counted. Differences arise only for complex forms, such as Indonesian *ke pada* 'to place > animate goal', which are separated by equal sign in the data matrix (*ke=pada*). Figure 10 (left) represents the semantic map for French where partially identical forms are counted as different (*ke=pada* is as different from *ke* 'to' as from *dari* 'from'). Figure 10 (right) is French again in a map where partially identical forms are counted as identical (*ke=pada* is as identical to *ke* as to *ke*). Up to now an intermediate solution has been applied which I think is

the most appropriate of the three, to count partially identical forms as intermediate (*ke=pada* is 50% "identical" with *ke*) (for French see Figure 3 above). While the choice of how to count identity does not make any major difference for Dimension 1, there are some modifications in Dimension 2. If partial identity is disregarded, the distance between animate goal and inanimate goal grows and animate goal is the extreme pole in Dimension 2. If partial identity is overrated, the distance between animate goal and inanimate goal gets smaller and companion is now the extreme pole in Dimension 2. This is because there are many languages such as Indonesian where local phrase markers for animate goal are complex and partially formally identical with inanimate goal markers. Companion markers are not less complex, but they exhibit less systematic relationships with other clusters. Note that what changes in Figure 10 is the distance between the clusters rather than the density of the clusters.

Obviously, counting all partial formal identities as 0.5 is not a sophisticated solution. Intuitively, complex forms are more closely related to longer parts and to parts with lower token frequency. Thus, intuitively, Italian *dietro=a* is more closely related to *dietro* than to *a*. There are certainly better ways of counting identity, but for the time being it seems to be a good solution to adopt an intermediate approach between the two extremes of disregarding and overrating partial identity.

Figure 10: Different ways of calculating the matrix: partially identical is different (left) and partially identical is identical (right)



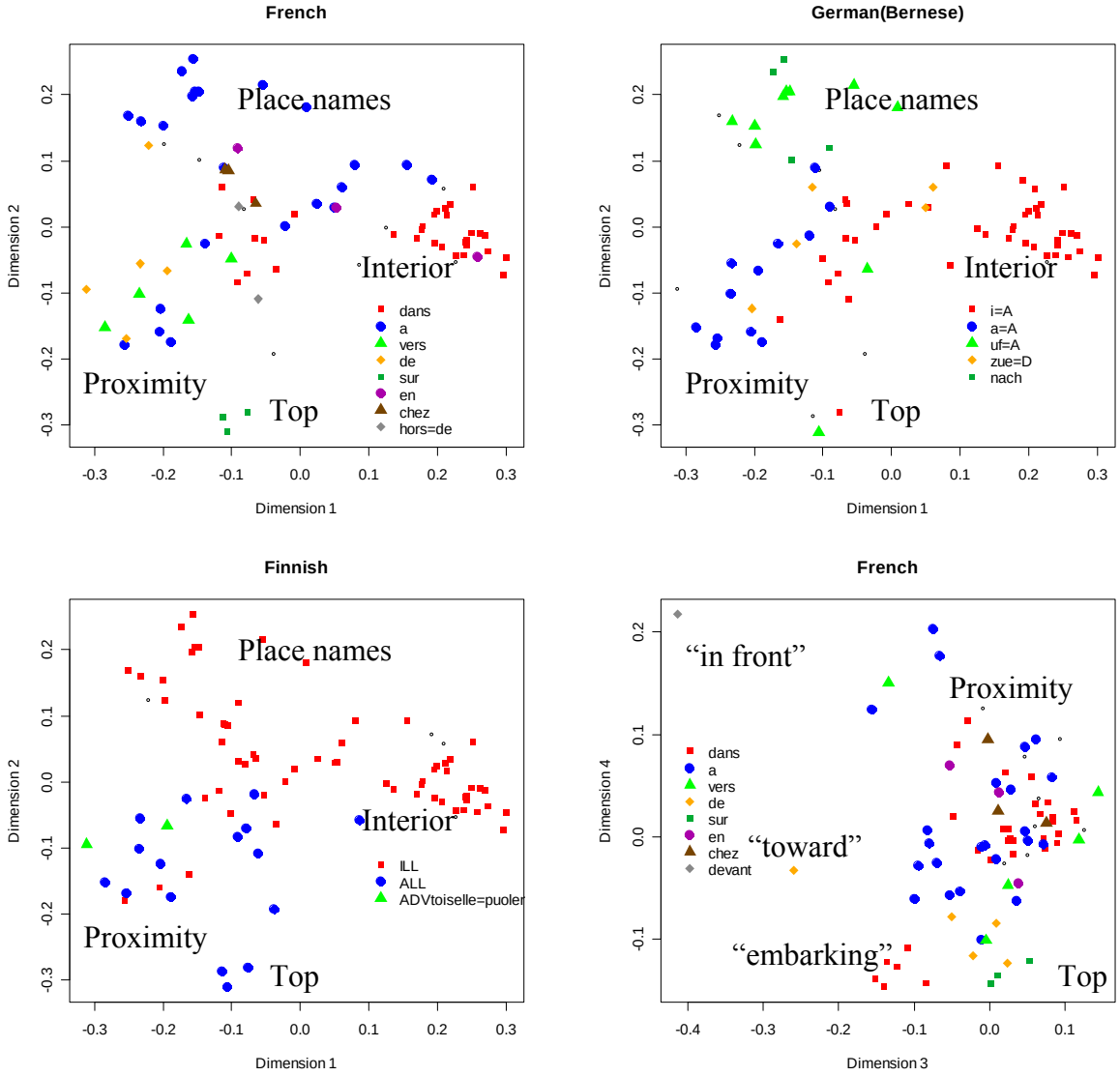
In the same way as resampling languages we can resample situations. For getting a better view on inanimate goal we can select only situations for inanimate goal in the database. The major dimensions discussed so far (role and animacy) will then disappear and the tendencies now emerging in the 153 language sample do not really amount to clusters because there are no evident clear trends in the data any more. What we get now in Figure 11 in Dimension 1 is an interior positive pole and a proximity negative pole. In Dimension 2 the negative pole is surface/top and the positive pole is place names. This dataset does not lend itself easily to clustering. There are many discontinuous categories, such as Bernese German *uf* with Accusative 'onto' which is also used for movement to place name. The three plots for French, Bernese German and Finnish show that place names can be combined with in a category with any of the three major localizations, with proximity in French (*a*), with surface/top in Bernese German (*uf*), and with interior in Finnish (*ILLative*).

Dimensions 3 and 4 are mapped only for French (Figure 11, bottom-right). Dimension 3 points out two particular situations in the negative pole: 1:33, the only situation where many

languages have “in front” (but not King James: *And all the city was gathered together* ***at*** *the door*), and 11:01, the only situation where many languages have “toward/approaching” (*And when they came nigh* ***to*** *Jerusalem...*). Dimension 4 has the poles proximity (positive pole) and top (negative pole) with some of the situations reordered. Embarking a boat goes now together with top (French text *dans*) rather than with interior as in Dimension 1 testifying to the fact that embarking a boat is intermediate between top and interior. Interestingly, higher dimensions can be better interpreted in this smaller dataset (84 instead of 190 situations). We have the situation here that the more general trends role and animacy are strong. Only if the sample of situations is chosen such that the strong major dimensions are removed can the contribution of the weaker dimensions with more restricted scope emerge.

If we remember from the discussion above that sampling of situations can be interpreted psychologically as focus of attention (activation in memory), a psychological interpretation of this finding is that semantic space can change considerably depending on different selected attention to particular sets of examples. Put differently, every semantic field or domain has its own semantic space. Some semantic distinctions will emerge only if attention is focused to a smaller set of activated items.

Figure 11: Semantic map of local goal markers (84 situations)



The purpose of this section has been to show that there is a multiplicity of similar possible semantic maps which can be built for a particular domain, depending on the languages sampled, the situations sampled, and the way of identifying and counting forms for calculating the similarity matrix. Further sources of variation not discussed in this section are the distance measure used for calculating the distance matrix and the visualization tool applied (different versions of multidimensional scaling, the neighbor-nets of Huson and Bryant 2006, etc.). However, the fact that there are many ways to build a semantic map does not mean that semantic space is vague or undetermined. Rather semantic space is more

powerful than assumed in traditional approaches to semantic maps. Semantic space is not stable, but dynamic. Croft (2001:109) makes a distinction between universal conceptual structure and language-specific semantic structure. This seems to me to be only a first step toward a model of dynamicity of semantic/conceptual space. The approach of Nosofsky and Palmeri (1997) suggests that psychological space is slightly different for every human being and changes over time with every new exemplar presented and with different degrees of attention paid to particular semantic dimensions. This line of reasoning leaves us with dynamic psychological semantic spaces in individuals and probabilistic spaces which are a kind of average psychological semantic spaces in certain populations (be it a language, a language family, a continent, or world-wide linguistic diversity).

A consequence is that there is no static universal semantic space. If we build semantic maps on the basis of large world-wide samples of languages we get averaged semantic space, where frequent semantic patterns clearly emerge and rare semantic patterns are hardly distinguishable from noise. As pointed out by Gil (2004:415) cross-linguistic semantic maps are "rapidly overwhelmed with an arbitrarily large number of arbitrarily specific 'small' functions". However, a very large number of specific 'small' functions can develop in dynamic semantic spaces emerging from constellations of exemplars with varying degrees of activation. Large numbers of local oppositions can emerge in multidimensional spaces, supported by language-particular formal differences, stretching space in various ways, all sensitive to similarity. Rather than a single universal semantic map there are as many psychological semantic spaces as human beings, all evolving through time, all very similar to each other, and all variations of each other without an underlying theme.

5 Conclusions and outlook

It has been argued in Section 2 that semantic maps have a theoretical foundation in similarity semantics and, as far as based on databases of contextually-embedded situations, on exemplar semantics. The semantic map approach in most of its facets is more empirical than many other approaches to semantics, but this does not imply necessarily that it is less theoretical. Whoever does not agree about the underlying theory of semantic maps should agree about the necessity of making explicit the theoretical foundations of the semantic map approach. Put differently, if we know that it works, we should also be interested in why it works. It is argued here that the empirical focus of the semantic map approach follows from the a priori unpredictable nature of similarity. Meaning emerges by way of semantic connections between exemplar situations based on similarity and the semantic network arising is constrained only by the unpredictable set of similarity relationships between any pair of exemplar situations, which differ, however, strongly in probability of occurrence. Semantic space is a probability space which can be modeled by statistical methods which need concrete databases as input.

It is also important to know what the underlying theory is because practical applications of the theory might require some assumptions which do little harm in practice but are problematic from a theoretical point if view. In my view, a fundamental difference between theory and practice is that the practical applications assume that the cross-linguistically identified analytic primitives (domains or situations) are identical when they are in fact only very similar. Practical applications of semantic maps are anti-relativistic, assuming complete identity of cross-linguistically identified functions. However, the underlying theory need not be anti-relativistic. Semantic maps work in practice to the extent that the cross-linguistically identified analytical primitives are less different in meaning than the ones compared within languages.

Semantic maps have certain "technical" or "optical" characteristics that are due to the method, not to the underlying theory, notably resolution and sharpness. All existing semantic maps have low resolution. Even if the "etic grid" is constructed by "teams of fieldworkers who have extensive experience of the languages they intend to investigate" (Levinson and Meira 2002:487), semantic maps are a very crude method due to the low resolution obtained. However, increasing the resolution makes only sense if the analytical primitives are sharp. A semantic map can show a sharp picture only if the analytical primitives are distinct from each other (the semantic differences between the domains/functions should be smaller than the cross-linguistic semantic difference). Contextually-embedded situations have the advantage that they tend to be sharper and so it is possible to have a larger number of pixels and thus to obtain a better resolution.

Functional equivalence means in practice translational equivalence. Rather than rejecting translation as method of obtaining maximally comparable data, it should be investigated what the concrete effects of practical translations are how they can distort semantic maps. An obvious assumption is that translation will entail a lower degree of structural diversity. However, in this paper parallel texts have been adduced to demonstrate exactly the opposite, the high amount of structural diversity in language use in local phrase markers which cannot be modeled on the basis of traditional implicational maps. It is certainly true that something is lost in translations, but parallel texts are very useful at least from a methodological point of view since they embody the ideal of translation equivalence in practice with all practical complications following from that.

Most approaches to semantics a priori focus on certain aspects of meaning, which they consider essential, and disregard all other aspects of meaning, which they consider non-essential, usually without explaining convincingly why exactly the semantic features chosen should be prioritized a priori (for a criticism of essentialist methodology in linguistics see

Altmann and Lehfeldt 1971:20-22 and Croft 2000:17, 26). Semantic maps are an empirical approach to semantic structure which has the potential to do away with many unnecessary a priori essentialist decisions. It is desirable to develop methods of building semantic maps from ever larger datasets with ever less preselection of data. The maps constructed in this paper are essentialist to the extent that they focus on a particular larger semantic domain (motion events) and have a particular definition of forms included in the database (local phrase markers). Parallel texts, the data source used in this paper, have the potential for even more radically non-essentialist semantic maps and if automatic morphological analysis (algorithmic morphology) once should make it possible to build semantic maps fully automatically from parallel texts.

Like in typological universals there is a dichotomy between implicational and statistical/probabilistic semantic maps. Probabilistic semantic maps, such as exemplified in this paper, have the advantage that they can be built on the basis of large datasets from language usage directly without previous abstraction of general semantic domains. They can be used to test whether a priori semantic dimensions are supported by language use. However, emerging dimensions of the automatic analysis are in need of the a priori postulated semantic dimensions for interpretation.

Semantic maps, like any other instrument of typological research, reflect the linguistic diversity they are based on, be it implicitly or explicitly in form of a database. As shown in Section 4, sampling is therefore equally relevant as in all other typological approaches and semantic maps can be used for areal typological research like other methods of quantitative typology. A semantic map based on African languages cannot be expected to be an ideal model to represent European languages and a map based on Indo-European languages will most accurately reflect Indo-European languages. It is desirable to have large samples of

languages and it is important to consider differences between various populations of languages (such as continents and large families).

Semantic maps are sensitive not only to sampling of languages but also to sampling of analytic primitives ("domains", "functions", dots on the map). Resulting maps are determined by the choice of analytical primitives as much as by the sample of languages. By choosing a certain set of analytic primitives Levinson and Meira (2003) have excluded a priori the two dimensions that emerge as the strongest tendencies in my investigation of local phrase markers (role and animacy). However, by doing so, they get a much better coverage of localization or topology, which is most clearly differentiated in the role residence which has been completely disregarded in this investigation based on motion events.

There is little doubt that having a large number of analytic primitives is desirable in semantic maps. Semantic maps are ideally based on large databases. Building semantic maps on the basis of large databases is not possible by hand. Fortunately, there are good statistical methods available, implemented in easy software tools (many of them open access), which is why I see no reason to draw semantic maps manually rather than having them built automatically. There is little hope that we will identify a single ideal method of building semantic maps rapidly, such as some linguists see it in the method presented by Croft and Poole (2008). Rather than declaring one and only one method as standard, we should start discussing the advantages and disadvantages of various methods which first requires that they can be easily replicated. To express this in the words of (Ogden and Richards 1923:101): "To discuss such questions in any other spirit than in which we decide between the merits of different weed-killers is to waste all our own time and possibly that of other people". There are many ways to represent the same data in slightly different semantic maps. There are multiple ways to calculate the distance matrix and there are different visualization tools. The underlying data are usually extremely diverse. Visualization always implies some amount of

data reduction. Semantic maps are a good tool for identifying the fundamental tendencies in the data. Usually they are no good tool to represent rare categories.

It has been argued in this paper that semantic maps rest on the isomorphism hypothesis which is an exception to de Saussure's *arbitraire du signe*. There are many unsolved questions which are related to this issue. If the isomorphism hypothesis is an exception to the *arbitraire du signe* there are maybe also other exceptions which might have an effect on semantic maps. For instance, unmarked forms need not necessarily be equally similar in meaning as identically marked forms. For short and even more so zero form, identity in form is more likely to be accidental than for longer forms. Another issue is whether similarity in meaning will always be reflected by identity in form. It is very well possible that some similar meanings will never happen to exhibit the same forms. Formal identity is conditioned to a large extent by diachronic pathways of semantic change and it may be that semantic change privileges certain forms of similarity which will then be overrepresented in semantic maps. To investigate such issues, we need more sophisticated models of similarity which probably requires a closer interaction of typology with psychology.

Another crucial issue is how to count identity of forms in building semantic maps. At present, most semantic maps are built on the basis of simple morphemes or categories, but it should also be possible to build maps on the basis of complex forms and constructions which are only partially identical to each other. In Section 4 it has been shown that it matters for probabilistic maps how identity is counted. The solution offered is that formal identity should be counted in different ways in order to assess the potential variation due to formal identity decisions.

Perhaps the most crucial issue for the semantic map approach in the future will be to better understand the nature of semantic space in its various manifestations. Understanding the relationship between psychological semantic space, averaged language-particular semantic

space, and averaged typological semantic space is indispensable for exploring the effects of categorization in particular languages.

Acknowledgements

I would like to thank an anonymous reviewer for many useful comments and Michael Cysouw for having introduced me into R and multidimensional scaling. This work would not have been possible without the help of many colleagues who supported me in getting access to Bible translations in many different languages. For the analysis of some texts I was supported by colleagues (especially Masayuki Onishi for Motuna and Søren Wichmann for Tlapanec). While writing this paper I was funded by the Swiss National Science Foundation (PP001-114840).

Abbreviations

ABL ablative, ACC accusative, CAUS causative, DAT dative, DEF definite article, DEST destinative, DIFF diffusive DIST distal, DS different subject, ELA elative, EMP emphasis, EXTR extreme, GEN genitive, ILL illative, IMMED immediate, IMP imperative, LOC locative, M masculine, MULTI multiple NEX nexus element, NM noun marker, OBJ object, PERF perfect, PL plural, POSS possessive affix, PP adposition, PROX proximal, PST past, PUNC punctual, PV preverb, RE realization, SG singular, SUBJ subject,

References

Altmann, Gabriel and Werner Lehfeldt. 1973. *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren*. München: Fink.

Berkeley, George. 1998 [1710 / 1734]. *A Treatise Concerning the Principles of Human Knowledge*, ed. by Jonathan Dancy. Oxford: Oxford University Press.

Borges, Jorge Luis. 1944 / 2005. *Ficciones*. Madrid: El libro de bolsillo.

Bowern, Claire. 2008. *Linguistic fieldwork: a practical guide*. Basingstoke: Palgrave Macmillan.

Bréal, Michel. 1897 / 1915. *Essai de sémantique*. Sixième édition. Paris: Hachette.

Clark, Eve V. 1993. *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Chang, Suk-Jin. 1984. *Korean*. Amsterdam: Benjamins.

Comrie, Bernard. 1986. Markedness, grammar, people, and the world. *Markedness*, ed. by F. R. Eckman and E. A. Moravcsik, 85-106. New York: Plenum.

Crazzolara, J. P. 1955. *A Study of the Acooli language. Grammar and vocabulary*. 2nd revised impression. London: Oxford University Press for International African Institute.

Croft, William. 2000. *Explaining Language Change. An evolutionary approach.* Harlow: Longman.

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

Croft, William and Keith T. Poole. 2008. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics.*

Croft, William. 2007. Exemplar Semantics. Draft. http://www.unm.edu/~wcroft/Papers/CSDL8-paper.pdf

Cysouw, Michael. 2007. Building semantic maps: the case of person marking. *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, ed. by Matti Miestamo and Bernhard Wälchli, 225-247. Berlin: de Gruyter.

Cysouw, Michael, Chris Biemann, and Matthias Ongyerth. 2007. Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *Sprachtypologie und Universalienforschung STUF* 60/2.158-171

Cysouw, Michael and Bernhard Wälchli (eds.). 2007. Parallel Texts. Using translational equivalents in linguistic typology. Theme issue in *Sprachtypologie & Universalienforschung STUF* 60/2.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13.257-292.

Ebert, Karen H. 1999. The UP – DOWN dimension in Rai grammar and mythology. *Himalayan Space. Cultural horizons and practices*, ed. by Balthasar Bickel and Martin Gaenszle, 105-131. Zürich: Völkerkundemuseum

Erdmann, Karl Otto. 1922. *Die Bedeutung des Wortes. Aufsätze aus dem Grenzgebiet der Sprachpsychologie und Logik.* Dritte Auflage. Leipzig: Haessel.

Fillmore, Charles J. 1971/75. *Santa Cruz lectures on Deixis.* Reproduced by the Indiana University Linguistics Club 1975. Bloomington, Indiana.

Ganenkov, D. S. 2002. Tipologija padežnix značenij: semantičeskaja zona prolativa. *Grammatikalizacija prostranstvennyx značenij. Issledovanija po teorii grammatiki 2*, ed. by Vladimir A. Plungian, 35-55. Moskva: Russkie slovari.

Gil, David. 2004. Riau Indonesian *sama*: Explorations in macrofunctionality. *Coordinating Constructions*, ed. by Martin Haspelmath, 371-424. Amsterdam: Benjamins.

Gilliéron, Jules. 1919. *La faillite de l'étymologie phonétique*. Étude sur la défectivité des verbes. La Neuveville: Beerstecher.

Goddard, Cliff. 1996. *Pitjantjatjara/Yakunytjatjara to English Dictionary*. Revised second edition. Alice Springs: IAD Press.

Goldsmith, John A. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27/2.153-198.

Goodman, Nelson. 1972. Seven strictures on similarity. *Problems and Projects*, Nelson Goodman, 437-447. New York: Bobbs-Merrill.

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

Hamming, Richard W. 1950. Error-detecting and error-correcting codes. *Bell System Technical Journal* 29/2.147-160.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. *The New Psychology of Language* 2, ed. by Michael Tomasello. 211-242. Mahwah, NJ: Lawrence Erlbaum.

Hjelmslev, Louis. 1961. *Prolegomena to a Theory of Language*. Translated by Francis J. Whitfield. Madison, Milwaukee: University of Wisconsin Press. [Revised version of *Omkring spogteoriens grundlæggelse* (1941). Copenhagen: Munksgaard]

Holton, Gary. 2003. *Tobelo*. München: Lincom.

Huson D. H. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23/2.254-267.

Kibrik, Aleksandr E. 1970. K tipologii prostranstvennix značenij (na materiale padežnyx sistem dagestanskix jazykov) [About the typology of spatial meanings (on the material of case systems in Dagestan languages)]. *Jazyk i čelovek. Sbornik statej pamjati prof. P. S. Kuznecova*, ed. by Roman Jakobson et al., 110-156. Moskva: Izdatel'stvo moskovskogo universiteta.

Kilby, D. 1981. On case markers. *Lingua* 54.101-133.

Konstanz Universals Archive http://typo.uni-konstanz.de/archive/intro/

Levinson, Stephen C. 2003. *Space in Language and Cognition. Exploration in cognitive diversity*. Cambridge: Cambridge University Press.

Levinson, Stephen and Sérgio Meira. 2003. 'Natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language* 79: 485-516.

Logan, G. D. 1988. Toward an instance theory of automatization. *Psychological Review* 95.492-527.

Malotki, Ekkehart. 1979. *Hopi-Raum. Eine sprachwissenschaftliche Analyse der Raumvorstellungen in der Hopi Sprache*. Tübingen: Narr.

Masica, Colin. 1976. *Defining a Linguistic Area*. South Asia. Chicago: University of Chicago Press.

Mauthner, Fritz. 1923 / 1982. *Beiträge zu einer Kritik der Sprache*. Erster Band: *Zur Sprache und zur Psychologie*. 2., vermehrte Auflage. Leipzig: Meiner / Frankfurt: Ullstein Materialien.

Mervis, Carolyn B. 1988. Early lexical development: theory and application. *The psychobiology of Down syndrome*, ed. by Lynn Nadel, 101-143. Cambridge, Mass.: MIT Press.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nosofsky, Robert M. and Thomas J. Palmeri. 1997. An exemplar-based random walk model of speeded classification. *Psychological Review* 104/2.266-300.

Nosofsky, Robert M. and Roger D. Stanton. 2006. Speeded old-new recognition model of multidimensional perceptual stimuli: modeling performance at the individual-participant and individual-item levels. *Journal of Experimental Psychology* 32/2.314-334.

Ogden, C. K. and I. A. Richards. 1923 / rev. 1927 / 2001. *The Meaning of Meaning. A study of the influence of language upon thought and of the science of symbolism.* / London: Routledge (I. A. Richards. Selected Works 1919-1938. Volume 2.)

Roberson, Debi, Jules Davidoff, and Nick Braisby. 1999. Similarity and categorisation: neuropsychological evidence for a dissociation in explicit categorisation tasks. *Cognition* 71.1-42.

Robert, Stéphane. 2006. Deictic space in Wolof: Discourse, syntax and the importance of absence. *Space in Language. Linguistic Systems and Cognitive Categories*, ed. by Maya Hickmann and Stéphane Robert, 155-174. Amsterdam: Benjamins.

Samarin, William J. 1966. *The Gbeya Language. Grammar, texts, and vocabularies*. University of California Publications in Linguistics 44. Berkeley: University of California Press.

Saussure, Ferdinand de. 1968. *Cours de linguistique générale. Édition critique par Rudolf Engler*. Tome 1. Wiesbaden: Harrassowitz.

Schopenhauer, Arthur. 1913. *Arthur Schopenhauers sämtliche Werke. Herausgegeben von Dr. Paul Deussen.* Neunter Band. *Arthur Schopenhauers handschriftlicher Nachlaß. Philosophische Vorlesungen. Im Auftrage und unter Mitwirkung von Paul Deussen zum ersten Mal vollständig herausgegeben von Franz Mockrauer.* Erste Hälfte. *Theorie des Erkennens*. München: R. Piper & Co.

Shepard, Roger N. 1987. Toward a universal law of generalization for psychological science. *Science* 237.1317-1323.

Vries, Lourens de. 2007. Some remarks on the use of Bible translations as parallel texts in linguistic research. *Sprachtypologie und Universalienforschung STUF* 60/2.148-157.

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

Wälchli, Bernhard. 2007. Advantages and disadvantages of using parallel texts in typological investigations. *Sprachtypologie und Universalienforschung STUF* 60/2.118-134.

Wälchli, Bernhard and Michael Cysouw. Forthcoming. Toward a semantic map of motion verbs. Explorative statistical methods applied to a cross-linguistic collection of contextually-embedded exemplars. Special issue of *Linguistics*: Lexical typology, ed. by Maria Koptjevskaja-Tamm and Martine Vanhove.

Wälchli, Bernhard and Fernando Zúñiga. 2006. Source-Goal (in)difference and the typology of motion events in the clause. *Sprachtypologie & Universalienforschung STUF* 59/3.284-303. (Theme issue *The Lexicon: Typological and Contrastive Perspectives*, ed. by Giannoula Giannoulopoulou and Torsten Leuschner).

WALS = Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.). 2005. The World Atlas of Language Structures. (Book with interactive CD-ROM). Oxford: Oxford University Press.

Xajdakov, Sajd M. and L. I. Žirkov. 1962. *Laksko-russkij slovar'*. Moskva: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

Appendix A

This appendix contains the Python (www.python.org) code that calculates the three distance matrices described in Section 4 and writes an R-code which generates plots of the MDS-analysis for all doculects in the input data table. To run this program, the libraries rpy (interaction of Python and R) and numpy (enabling Python to use matrices of the kind R uses them) not contained in the basic Python package must be installed and the interaction of Python and R works slightly differently with different versions of Python and R and on different platforms. As the program is written, the input text file must be saved in ANSI with the fields separated by tabs or spaces. The first two columns on the left contain data labels (identification of situations) and the first row contains language labels. The first row with the language labels begins directly with the labels (thus, this row has two fields less than all other rows). No cells may be empty and no cells may contain spaces. The following strings are treated as non-attested: "NA" (upper case only), "?", and "_".

Places in the program which must be adapted on every computer are indicated by ***. The name of the input file is defined in the program.

The output files have the same name as the input file plus the following extensions: "langlist.txt" the R code, "fuerR.txt" the input data for R, "rownames.txt", "colnames.txt" the situation labels and doculect labels for R, "wholemix.txt", "whole.txt", "wholeor.txt" the three distance matrices for R. If there are any files by the same names, the program replaces these files.

This program is free software and comes without any guarantee.

INSERT LINK TO PYTHON PROGRAM HERE