Open access • Journal Article • DOI:10.1515/LING-2012-0021

Lexical typology through similarity semantics: Toward a semantic map of motion verbs — Source link

Bernhard Wälchli, Michael Cysouw

Institutions: Max Planck Society

Published on: 18 May 2012 - Linguistics (Walter de Gruyter GmbH & Co. KG)

Topics: Semantic similarity, Lexical semantics, Semantics and Similarity (psychology)

Lexical typology through similarity semantics: Toward a semantic map of motion verbs*

BERNHARD WÄLCHLI AND MICHAEL CYSOUW

Abstract

This paper discusses a multidimensional probabilistic semantic map of lexical motion verb stems based on data collected from parallel texts (viz. translations of the Gospel according to Mark) for 100 languages from all continents. The crosslinguistic diversity of lexical semantics in motion verbs is illustrated in detail for the domain of 'go', 'come', and 'arrive' type contexts. It is argued that the theoretical bases underlying probabilistic semantic maps from exemplar data are the isomorphism hypothesis (given any two meanings and their corresponding forms in any particular language, more similar meanings are more likely to be expressed by the same form in any language), similarity semantics (similarity is more basic than identity), and exemplar semantics (exemplar meaning is more fundamental than abstract concepts).

1. Introduction

This paper explores how lexical semantics can be approached by direct crosslinguistic comparison of contextually embedded examples.[1] The basic problem of lexical comparison across many languages is that the ranges of meanings of language-particular lexemes are highly variable, possibly too variable to be directly comparable. As a solution to this problem, we will use a large set of concrete contexts and investigate which lexeme is used in each context. We then compare the distribution of lexemes over these contexts across languages to get an impression of the similarities and differences between the lexicalization of different languages. In the terms used by Koptjevskaja-Tamm (2008: 11) this is the denotation rather than the sense approach (see Koptjevskaja-Tamm 2008 for an extensive review of other solutions to the same problem).

The tool we propose to use is a probabilistic semantic map, illustrated in this paper by the domain of motion verbs in a massively parallel text. Semantic maps are an empirical approach to semantics which usually rest on massive

Linguistics 50–3 (2012), 671–710
DOI 10.1515/ling-2012-0021
0024–3949/12/0050–0671

crosslinguistic comparison. In Section 2 we discuss the three basic notions of crosslinguistic comparison used in this paper: *doculects* (our replacement for "languages"), *language-particular form classes* (our replacement for "categories") and *contextually embedded situations* (our replacement for "functions"). The crucial advantage of parallel texts is that we can compare languages on the level of individual contextually embedded examples rather than on the level of abstract systems.

The theoretical foundation of similarity semantics is discussed in Section 3. All semantic maps, it is argued, rest on the Isomorphism Hypothesis that identity in form reflects similarity in meaning (Haiman 1985: 19). This makes semantic maps necessarily an indirect approach to semantics; semantics is accessed by way of form. In our view, this isomorphism is not absolute but probabilistic. The configuration of the elements in the semantic map reflects the probability of them being expressed by identical or different categories across the sample of languages on which the map is based.

Since we find a lot of diversity even in a restricted search space, the general question arises as to how regularity relates to diversity and what this means for crosslinguistic analysis. In Section 4 we argue that high amount of diversity and high amount of regularity do not exclude each other, but rather ask for robust tools of analysis which can account for idiosyncrasies and general trends at the same time. It is argued that probabilistic semantic maps are an appropriate tool in this endeavor.

Probabilistic semantic maps differ from traditional implicational maps (Haspelmath 2003) in that they can cope with large sets of elements and with messy data. This means that we need not first posit abstract homogeneous functions, but can build semantic maps directly on exemplar data. Probabilistic semantic maps have the additional advantage that they can be built automatically by means of standard techniques of statistical analysis. In a two step procedure we first calculate a distance matrix of all elements to be displayed and in a second step the elements are arranged on a number of dimensions by the method of multidimensional scaling (MDS) (Section 5). Sets of two dimensions can then be visualized as a two-dimensional map.

Section 6 deals with the concrete data set on which probabilistic semantic maps are illustrated: motion verbs in 360 situations across translations of the Gospel according to Mark (henceforth: Mark) across 101 doculects. As many other domains of lexical typology, motion event encoding is highly complex. The method used allows us to approach this complexity in terms of the number of dimensions needed to capture the most general trends in the crosslinguistic data. The sample of doculects used is strongly biased toward European languages and the sample of motion event situations considered is far from covering the whole diversity of motion events. However, even with such a restricted search space with highly limited diversity we clearly need more than ten

dimensions to account for even the most general trends in the data. It is shown that probabilistic maps are a good tool for capturing the general crosslinguistic trends in the database, but are not appropriate for accounting for the many rare categorization patterns in the sample.

Since it is not possible to consider all aspects of motion events at the same time, we focus on two particular dimensions in Section 7. MDS allows us to do so: we can arrange the situations on two dimensions and disregard all the other dimensions. Section 7 discusses the diversity of category types found in these dimensions and offers some directions as to how synchronic probability maps can also be useful for diachronic analysis.

Since diachrony plays a major role in the semantic map discussion, although it is not the major focus of our paper, Appendix A addresses the question as to how probabilistic semantic maps relate to diachronic semantic maps in more general terms.

2. Method: primary data typology

The basic problem of lexical typology (like all crosslinguistic comparison) is to find a way to compare like with like (cf. Koptjevskaja-Tamm 2008: 45). To achieve this, we posit the following three kinds of entities as basic units of analysis: doculects, language-particular form classes, and contextually embedded situations:

Doculects are our replacement for the notion of language. A doculect is any documented language variety, be it as raw data (e.g., a sound file), primary data (e.g., a transcribed text or wordlist), or secondary data (e.g., a glossed text or a grammatical description) of whatever size.[2] *Language-particular form classes*, or language-particular categories, are lexemes, morphemes, or constructions used recurrently in utterances of a particular doculect. Such elements are used in *contextually embedded situations*.[3] Instead of assuming abstract functional domains (e.g., the "functions" of Haspelmath 2003), we use concrete instantiations of particular functions as determined by a given context for typological comparison. In our approach, abstract functional domains are not homogeneous entities, but clusters of exemplars.[4] Using these definitions we can summarize our approach to lexical typology in one sentence by saying that lexical comparison is the comparison of the distribution of lexical form classes from different doculects across a sample of contextually embedded situations.[5]

For the collection of contextually embedded expressions we use *massively parallel texts* (Cysouw and Wälchli 2007) in the form of translations of 360 situations from the Gospel according to Mark describing motion events (see Section 6 for more details on the data). Such translations of the same text make it possible to compile a large database of crosslinguistically comparable

expressions in a large number of diverse languages. This comes at the cost of some idiomaticity due to translation (see Wälchli 2007 for a discussion of using translations of Bible texts, and see Section 6 for an example of the kind of problems that arise). However, using translations is actually nothing else than the practical implementation of the abstract idea of translational equivalence, which is pervasive in functional linguistics.

Parallel texts are one of several data sources in typological research based on primary data. *Primary data typology* is a cover term for all typological data collection based on primary sources (rather than on descriptions or other high-level analyses) and on exemplars (rather than abstractions or other complex categories). Other primary data sources that can be used are translational questionnaires (e.g., Dahl 1985; Ricca 1993), nonverbal questionnaires in the form of series of pictures or video clips (e.g., Bowerman and Pederson 1992; Levinson and Meira 2003), retold stories (e.g., Chafe 1980; Bickel 2003), and original texts (e.g., Myhill 1992; Güldemann 2008). Retold stories and original texts are the more naturalistic form of language use, but it is more difficult to use them for language comparison, because the functional parallelism is more difficult to establish. Controlled questionnaires elicited under experimental conditions in the natural environment of the speakers of a language are of course to be preferred over literary translations. However, such data is normally difficult to obtain for a large number of structurally diverse languages.

3. Theoretical foundation: similarity semantics and exemplar semantics

This paper advocates a method of comparing languages on the level of exemplars rather than on the level of abstract concepts. Its theoretical basis is *similarity semantics* and *exemplar semantics*. The definitions given here are our own, but related notions are used by many philosophers, psychologists and linguists. Similarity semantics is a cover term for all approaches to semantics where similarity is considered to be a more basic notion than identity. Exemplar semantics is a cover term for all approaches to semantics where exemplar meaning is considered more fundamental than the meaning of abstract concepts (see also Croft 2007). This section discusses similarity semantics and exemplar semantics in a theoretical semantic context.

Approaches to semantics differ as to whether *identity* or *similarity* is considered a more basic notion. In many theoretical approaches to meaning, the primacy of identity is not made explicit; it is simply taken for granted.[6] Marty (1908: 407), for example, distinguishes two kinds of similarities, both derived from identity: partial identity of complexes ("teilweise Gleichheit von Zusammengesetztem") and close species of the same genus ("nahestehende Spezies derselben Gattung"). In contrast, a fervent advocate of the primacy of similar-

ity was Fritz Mauthner: "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 . . ." (Mauthner 1982 [1923]: 469).[7] For Mauthner, similarity judgments are a necessary precondition for language to function. Conceptualization is possible only because the senses are not infinitely sharp and humans therefore treat strong similarity as identity. If pairs of meanings are chosen such that the difference between them constantly decreases, at some point the semantic difference becomes too small to be perceivable. The two meanings will then be *indistinguishable*, but this does not necessarily entail that they are identical; they might just represent a pair of two extremely similar meanings.[8]

The meaning of a form class, such as words (e.g., *walk*), morphemes (e.g., *-ness*), and constructions (e.g., a phrase 'determiner + noun'), can be approached in two different ways. The meaning can be considered to denote an abstract concept or it can be considered to be a range of individual meanings of exemplars. Using exemplary models of meaning has a long tradition. It has an early philosophical predecessor in George Berkeley who rejected John Locke's notion of concept ("idea") as an abstract entity: "an idea which, considered in itself, is particular, becomes general by being made to represent or stand for all other particular ideas of the *same sort* . . . 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 1998 [1710]: 94 [1710/1734: Section 11]). Similarly, for Ogden and Richards (1966 [1923]: 101) concepts are only "conveniences in description, not necessities in the structure of things". Clearly, exemplar meanings expressed by a single form class (e.g., all instances of English *walk*) tend to be similar. This is formulated in Haiman's *Isomorphism Hypothesis*: "Different forms will always entail a difference in communicative function. Conversely, recurrent identity of form between different grammatical categories will always reflect some perceived similarity in communicative function" (Haiman 1985: 19). We would like to change this hypothesis slightly. First, Haiman's restriction to "grammatical categories" is not necessary, and, second, the isomorphism hypothesis should be formulated more probabilistically: given any two meanings and their corresponding forms in any particular language, more similar meanings are more likely to be expressed by the same form.

Many semantic theories view meaning as *compositional*. In this view, meanings of words can be decomposed into more basic elements and can be exactly paraphrased by using such basic elements. In a way, compositional approaches to meaning assume that lexical semantics is of the same basic nature as syntax.

The basic elements of lexical decomposition are reminiscent of words, and the relations between these "word-like" semantic components are in some way syntactic. In syntactic theory, the mainstream view that constituents are more basic than constructions has recently been challenged by various versions of construction grammar, notably Croft's (2001) *Radical Construction Grammar*. In semantics the same question arises about what is more basic: abstract semantic components or the individual contextually embedded situational meanings. Exemplar semantics, in a construction grammar sense, proposes to take individual situational meanings as basic. This means in practice that each use of a lexeme is taken as a nonanalyzable unit. There is no attempt made to subdivide lexemes into smaller units. Generalizations are made empirically over actual occurrences, not by subdividing occurrences into smaller parts.[9]

Further, approaches to meaning can be direct and *absolute* or indirect and *relational*, depending on whether meanings are considered to be entities of their own definition, or whether they are defined in relation to other entities. In an absolute approach to semantics a "meaning" is considered to exist in some way, either as specific features or in the form of abstract concepts. In contrast, similarity semantics is purely relational; no statements are made to specify a certain meaning in concrete terms. It is only considered how similar a meaning is to other meanings. The similarity relationships of all exemplars constitute the semantic space. In this respect, similarity semantics is close to structuralist semantics (de Saussure 1968, Hjelmslev 1961: 51–54 [1943: 48–50]) with the difference that similarity semantics does not attribute the same relevance to establishing boundaries for strictly partitioning the amorphous "thought mass".

To summarize, the approach taken here assumes that similarity is more basic than identity in semantics. Our approach does not rely on any notion of semantic concepts, but instead assumes that form classes correspond to ranges of highly similar situational meanings, without drawing strict boundaries between such ranges. It is assumed that situational contextually embedded meanings are the real analytic primitives, and that there is no need to decompose them further into underlying semantic primes. Semantics then is constituted by the similarity relationships between exemplars. Meanings can be compared across languages by mapping them onto a crosslinguistic model of semantic space. Such a crosslinguistic semantic space can be obtained by taking an average over many language-particular semantic spaces. In effect, this implies that a crosslinguistic semantic space is an alternative to a semantic metalanguage.

4. Requirements for lexical-semantic analysis

In this section we will formulate general requirements for lexical-semantic analyses and we will explain why MDS analyses of massively parallel text data are a good tool for meeting many of these requirements at the same time.

The first requirement for lexical-semantic analyses is that they must be typological in the sense of *massively crosslinguistic*. It is intrinsic in the very nature of categorization that it simplifies semantic space considerably in a particular language. What categorization is like in a domain can therefore only be assessed if many languages are compared. In order to make description effective it must be known in what ways categories of lexical items can vary crosslinguistically. This does not mean that a dataset must cover the whole range of possible variation; if there is a high amount of diversity and/or if there are many rare categories, universality can never be reached. What we are mainly interested in is therefore not rare spectacular categorization patterns restricted to a few exotic languages but rather the amount of possible variation between frequent types of crosslinguistically highly similar categories.

The second requirement is to assess the *amount of diversity* in crosslinguistic categorization. If diversity is highly restricted in clearly predictable ways, we can use simple tools for description, such as binary features or simple checklists. It will be sufficient to identify a few major parameters and to classify languages into a selected number of types. Theory will then have to focus on the constraints that restrict diversity. However, if diversity in categorization patterns is high and categorization patterns are rather unpredictable, we need more sophisticated tools of analysis that can cope with many, and partly unexpected, dimensions of variation simultaneously while at the same time assessing the relative relevance of all the individual aspects of variation involved. What makes our approach to lexical typology different from many other approaches is that diversity comes before semantic analysis. We want to know first how diverse categorization is expected to be. Only once this step has been taken, can we choose the appropriate tools of semantic analysis.

The third requirement is to assess the *amount of regularity* in crosslinguistic categorization. Diversity and regularity are less strongly connected than is commonly believed. Regularity, as we understand it, is the major trends in the data. A dataset can be highly diverse and at the same time exhibit strong general trends. If there is great diversity without regularities, distribution is random and there is no point in developing sophisticated theories and methods of description. The greater and stronger the regularities are, the more analysis is needed.

The fourth requirement is to check to what extent the observable variation in categorization patterns is *semantic or asemantic*. In lexical-typological approaches it is usually taken for granted that major crosslinguistic differences in lexicalization patterns have semantic correlates. However, it might be the case that there are other factors involved and thus it should at least be assessed whether the stronger trends in the data have straightforward semantic correlates.

The fifth requirement is to assess the relevance of *usage-related factors*, such as *frequency* and *discourse structure*. Lexical semantics does not occur in

isolation but is shaped by language use in natural discourse like any other property of language.

The sixth requirement is to account for *diachronic developments*. There should be descriptive tools to model language change and the theory should account for what exactly changes in diachronic developments and to what extent these developments are crosslinguistically recurrent.

The seventh requirement is to provide tools for the *identification of attested categorization types*. It should be possible to state why two categories in two languages are very similar or in what ways the two categories differ and whether or not certain categories are sensitive to a semantic dimension reflecting a general trend in the data.

5. Practice: probabilistic semantic maps

Our database of motion events from Mark is too large to allow for a manual extraction of generalizations. General tendencies have to be extracted by means of well-established explorative statistical methods with as little data reduction as possible. We will use a variant of the semantic map methodology in the form multidimensional scaling (MDS) here. Using such a statistical method implies that the same procedures will be applied to all data in the database. There are no exceptions; no "irregular" data to be "explained away". If a particular language shows an idiosyncratic pattern, this will simply be downplayed among the many more common patterns.

Semantic maps have become an important tool of typological research during recent years. Major contributions include Kemmer (1993), Haspelmath (1997), van der Auwera and Plungian (1998), and Croft (2001). Further references to earlier work can be found in van der Auwera and Plungian (1998: 86–87), and more recent discussions can be found in the papers in Malchukov et al. (2010) and in the commentary to Croft and Poole (2008). A general survey of this approach is given by Haspelmath (2003). According to Haspelmath (2003: 213), semantic maps are a "method for describing and illuminating the patterns of multifunctionality of grammatical morphemes that does not imply a commitment to a particular choice among monosemic and polysemic analyses". In our case, the "multifunctionality" can better be described as "multicontextuality" as each contextual situation is seen as a separate function.

The goal of the semantic map approach is to investigate the structure among functions through the distribution of form classes (lexemes, morphemes, constructions) from many languages over these functions. Although traditionally used for grammatical functions, the semantic map approach in general is equally well suited for lexical meanings (cf. van der Auwera and Plungian 1998: 86; François 2008). In order to get a general perspective on the complex

and diverse relationships between meanings in language use, a large and diverse sample of languages must be considered. The aim of a semantic map is to represent the functions/meanings in such a way that the same form classes across all languages are represented as compactly as possible. These lexicalization patterns determine the underlying "map", (called "cognitive space" by Croft 2001: 92).[10]

We do not believe that the underlying structure of a semantic map necessarily reflects mental representations of meaning (as argued by Croft 2001: 92 by using the term "conceptual" or "cognitive" space). In our view, the underlying structure is a *probability space*. The closer two contextually embedded 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. A probability space is accurate to the extent that it predicts crosslinguistically recurrent tendencies in the categorization of form classes.[11] A probability space is largely determined by universal tendencies in cognition and general discourse conditions, and partly by historical coincidences. It is a model about synchrony, not diachrony. For a discussion of the relationship to diachronic semantic maps see Appendix A.

More importantly, the probability space is a nonverbal tool for describing and comparing meaning across languages. It is a continuous and empirically obtained alternative to discrete semantic metalanguages, like feature-based componential approaches. The difference in distribution of form classes across the contexts in the probability space represents the difference in meaning.[12]

As argued in Cysouw (2007, 2010), three steps are necessary to produce a semantic map of crosslinguistic variation: (a) a set of analytical primitives as the basis for crosslinguistic comparison, (b) a set of empirical relations between all pairs of primitives, and (c) a (visual) technique to help in the interpretation of any structure among these relations. This holds both for traditional implicational semantic maps and the probabilistic semantic maps. Table 1 compares the processing chain for traditional implicational semantic maps (Haspelmath 2003) with the semantic maps as constructed in this paper.

Table 1. *Processing chain in building semantic maps*

	Analytical primitives	Expression of primitives	Relations between pairs of primitives	Tool for interpretation
Implicational semantic maps	Idealized functions	Abstract translational equivalents	Identically coded in at least one language	Lines between adjacent functions
Semantic maps in this paper	Contextually embedded situations	Translations from parallel corpora	Hamming distance	Multidimensional scaling (MDS)

In implicational maps there are a small number of idealized functions that do not take into account the large amount of domain internal diversity of general abstract functional labels. For instance, Haspelmath's (2003: 215) map of the boundaries of French *à* includes both the functions "predicative possessor" and "experiencer". Predicative possessor is included because of examples such as *Ce chien est à moi* 'This dog is mine'. However, Predicative possession is excluded in the map for English *to*, obviously because of *This dog is mine* even though you could say also *This dog belongs to me*. No reference is made to the many cases of predicative possession in French where *à* is impossible such as *J'ai un chien* 'I have got a dog'. For experiencer the English example *This seems outrageous to me* is given which translates to French without *à* (*Ceci me semble indigne*) in at least one of its readings. It is not difficult to imagine many further French experiencer contexts without *à*. However, the function "purpose" is excluded from the *à* area even though *à* occurs in *à cause de* 'because'.

That implicational semantic maps rely on a very abstract notion of translational equivalence is usually not made explicit in the literature (it is not mentioned in Haspelmath [2003], for example). However, it follows from the fact that many of the examples adduced are ad hoc translations of selected examples which seem particularly suitable for characterizing a functional domain. In our approach we use actual translations in real texts instead.

In the current case we use the contextually embedded situations as the set of analytical primitives. That is, the predicates of motion as described in each of the 360 clauses chosen from Mark are the "points" in our semantic space. To build a semantic map among these we investigate the relation between those points on the basis of how they are expressed in the doculects of the sample. For traditional semantic maps, possible relations between the points are restricted to a present/absent dichotomy, with two primitives being either "attested as combined into the meaning of a language-particular category" or "unattested as such". However, it seems to be more useful to replace this binary opposition by a gradual notion (see Cysouw 2007 for a detailed argumentation of why this makes more sense). In practice, this means that for each pair of "points" (i.e., each pair of clauses chosen from Mark) we establish how similar (or distant) they are. The simplest measure for this similarity between two such points is the Hamming distance, which is the fraction of languages in our sample for which the relevant predicates of motion are lexicalized differently. Many different ways to calculate the distance matrix are possible (a discussion of which falls outside of the scope of this paper). The Hamming distance is the simplest one.

Let us illustrate this approach with the database used in this paper. Table 2 displays a small portion of the database. A distance matrix between situations is computed by using the (relative) Hamming distance as a distance measure.

For any pair of situations the number of differences attested in encodings is divided by the total number of the number of doculects in which both encodings are attested, which results in a distance matrix of 360 × 360 cells with "1" for maximally different and "0" for maximally similar. To exemplify this only for the seven languages given in Table 2, the situations Mark 6:29 and 9:13, for example, have a distance of 2/7 because two of seven pairs are different (Mapudungun *amu* : *küpa* and Spanish [Senc.] *ir* : *venir*). In situation 6:48b, one cell is not attested (NA) therefore we divide by six, since we cannot decide whether NA is the same or different. Together with 6:29, 3 of 6 pairs are different, hence the dissimilarity value is 0.5. The example also illustrates that seven doculects would be too little data to establish a clear picture, given that the encoding across the situations sampled is highly diverse crosslinguistically.

Table 2. *Extract from the underlying database*

Situations	English	French	Hungarian	Mapudungun	Mari	Spanish (Senc.)	Spanish (RV)
1:31	come	s=approcher	megy	fülkon	mijaš	ir	llegar
4:4	come	venir	jön	aku	čongeštaš	venir	venir
5:1	come	arriver	ér	puw	mijaš	llegar	venir
6:29	come	venir	jön	amu	tolaš	ir	venir
6:48b	come	se=diriger	indul	NA	mijaš	ir	venir
9:13	come	venir	jön	küpa	tolaš	venir	venir
9:33	come	arriver	ér	puw	tolaš	llegar	llegar
10:21	go	aller	megy	amu	kajaš	ir	ir
12:14	go	partir	megy	amu	kajaš	irse	partirse
14:3	come	entrer	lép	aku	puraš	llegar	venir

Performing these calculations for all pairs of analytical primitives results in a fully specified dissimilarity matrix specifying the relations between the points. The last step, visualization, is strictly speaking not necessary for modeling the crosslinguistic variation. All information is already present in the (dis)similarity matrix. However, visualization is needed because the human eye (and brain) is incapable of seeing structure in (large) dissimilarity matrices. For example, with 360 clauses, there are 360 × 359/2 = 64,620 different pairs for each of which a similarity is specified in the matrix. Such a high number of relations is simply too large to be interpretable by a human researcher without any help. In this paper, we will use multidimensional scaling (MDS) as a visualization tool. This is not the only possible choice, though we find MDS suitable for the current data.[13] Among various techniques of data mining MDS has the advantage that it arranges the analytic primitives along several dimensions which are scales. The poles of these dimensions can be interpreted.

The basic idea of MDS is that a large set of dissimilarities is organized into a spatial approximation. All points are ordered along dimensions that approximate the original dissimilarities. Roughly spoken, a first linear dimension is computed to order the points on a linear scale in such a way that the ordering fits as good as possible to the empirical dissimilarities. Of course, the fit will for the most part not be perfect, so subsequent dimensions are used to refine the spatial approximation. For n points maximally $n - 1$ dimensions are needed. However, in most cases much less dimensions suffice to model the empirical dissimilarities. The dimensions themselves are abstract mathematical constructs, only aimed at maximizing the fit. They are centered around zero, though the values on these dimensions do not really have an inherent meaning. Of interest is the relative ordering of the point on the dimensions, and whether the dimensions have any interpretative correlate.

The MDS analysis does not tell us what the descriptive meaning of a group of situations is. However, it tells us that wherever there is a cluster it is likely to have some descriptive correlate. Furthermore, if there is some intermediate cluster between two poles of clusters it is likely to have an intermediate meaning between the meanings of the pole clusters. The configuration obtained by the MDS analysis is thus a heuristic tool how to proceed in the descriptive analysis of the lexical domain covered by the data. We will attempt to interpret descriptively the poles of each dimension, starting with the lower dimensions which contain more information (recall that higher dimensions are only adjustments in addition to the lower dimensions). Subsequently we will proceed to interpret more subtle differences corresponding to the configuration of similar categories in different languages.

Interpretation is not possible without labeling of some sort and this is where a kind of metalanguage comes in through the backdoor in our approach. However, these labels — or descriptive meanings — are no semantic primes but rather have a similar status to the lexical part of glosses in grammatical analyses in reference grammars. The labels are not precise semantic definitions, but rather indicate that there is a certain range of meaning that is recurrently expressed by a number of categories in different languages and is therefore likely to represent a *crosslinguistic semantic category type*. The labels are necessarily fuzzy since they abstract away from the contextual meaning of many different situations and they bundle situational meanings to the extent that they are recurrent crosslinguistically. In our approach it is important that such interpretations are not made by introspection, but only where they are supported by a cluster of contexts based on similarity. In concrete terms, the poles of the dimensions can be interpreted either by considering the meaning shared by the situations with the most extreme values in a dimension or by considering the categories in the languages of the sample that cluster at a particular pole. We could arbitrarily label the clusters by category names in any language that

fits the cluster best, but for convenience we try to come up with English labels wherever possible.

The method is based on the assumption that descriptive semantic meanings are represented by clusters in the MDS analysis and the result confirms this assumption to a high degree since a large portion of clusters can be interpreted descriptively. All lower dimensions can be interpreted descriptively at least in one pole. Mismatches can be of two kinds. On the one hand, clusters can occur that appear impossible to be interpreted descriptively, this happens especially in higher dimensions, which is an artifact of the method. On the other hand, there can be categories which do not correspond to clusters at any dimension if the descriptive meaning of the category does not follow a general recurrent trend in the data considered. The method thus helps us to tell apart common category types from rare categories.

It is important to realize that any model obtained in this way will never represent the complete linguistic variation. Every linguistic domain can be represented by semantic maps in many slightly different ways and the resulting semantic maps will always depend to a certain extent on choices made by the investigator. Aside from the language material considered, necessary choices are the analytical primitives used as objects of the map, the method to compute the distances between the objects, and the choice of the visualization tool. More specifically, the analytical primitives are always a selected sample of functions or situations and will never reflect the full range of meanings that might be relevant. Sampling of analytical primitives is thus an issue that is at least as important as sampling of languages for the semantic map approach. Further, the relationship between different stages in the processing chain from data to visualization is never one-to-one. First, there are always different ways of coding the expression of analytic primitives (see, e.g., Wälchli 2010 on how semantic maps can change if markers sharing formal elements are treated as identical, partially identical, or different). Second, different distance measures prioritize different kinds of similarities and distances. Finally, visualization tools or clustering algorithms will represent only a part of the information in the distance matrix, and different techniques will result in slightly different maps. The choice between the various maps is not one between right and wrong, but one between suitable or unsuitable for a particular goal.

6. Data: a multidimensional semantic map of motion verb stems

The semantic map constructed in this paper rests on a database of motion verbs in 360 situations in Mark in 101 texts from 100 languages (Spanish is sampled twice, 16th century *Reina Valera Antigua* Spanish and Modern *en lenguaje sencillo* Spanish because we will need the two varieties of Spanish to illustrate a diachronic development).[14] The sample of languages (see appendix) contains

languages from all continents, but it is biased geographically toward European languages, and genealogically toward Indo-European, Uralic, and Austronesian languages. There are 4.6% of empty cells in the database, due to clauses missing in the translations. All in all, the database has 34,680 entries.[15] Only verb stems have been included in our current analysis. Inflectional affixes and verb particles (Talmy's [2000] satellites) have been disregarded for the purpose of this paper.[16] However, it is not clear in Talmy's framework what should be done with reflexive/middle verbs. In the approach pursued here reflexive/middle lexicalizations are considered to be different verbs. This is why, for instance, Spanish *irse* 'depart' is coded as a verb stem of its own different from *ir* 'go'. What is most important here is that a clear decision must be made on what are considered to be different verbs or same verbs, and the decision on what is considered to be a verb stem must be done by the researcher compiling the database. The reason why we decided to take reflexives/middles as a separate category is that in many languages "go self" is lexically associated with the meaning 'depart'. Disregarding reflexives/middle markers would entail a weaker differentiation of the 'depart' domain.

Only one verb per clause has been coded even in languages that regularly use multiverb constructions (like serial verbs, lexicalized converb constructions, or root serialization). Further, in all cases of multiverb constructions the more lexical of the verbs has been coded in the database (e.g., 'run' in a combination 'run'+'come', or 'take' in a combination 'take'+'go').[17] The numerous difficult language-specific decisions that were taken in the coding of the data will not be further discussed here.

Technically, the input for MDS is the distance matrix discussed above and the output is a matrix with values ranging from −1 to 1 for every data point in every dimension. Table 3 gives a small portion of the matrix for the seven situations with the lowest values for Dimension 1. As expected, the situations with the lowest and highest values in a dimension are semantically closely related. In Table 3 it can be seen that the negative pole of Dimension 1 has the meaning 'come'. Since it is not very practical to list all values numerically, the values are rather plotted than listed below (Figure 1).

Table 3. *A small portion of the MDS output with lowest values in Dimension 1*

Situations	*King James* English	Dim. 1	Dim. 2	Dim. 3	Dim. 4
4:4	and the fowls of the air came	−0.516	−0.008	−0.029	−0.014
91:3	Elias is indeed come	−0.511	0.0594	−0.05	−0.006
13:6	many shall come in my name	−0.501	0.0912	−0.024	0.0009
15:36b	let us see whether Elias will come	−0.5	0.1065	−0.027	−0.001
3:31	there came then his brethren	−0.496	0.0786	−0.04	0.0053
12:42	there came a certain poor widow	−0.495	0.1049	−0.029	−0.005
1:24	art thou come to destroy us?	−0.495	0.0988	−0.029	−0.005

From the MDS analysis of all this data it turns out that the movement domain consists of very many interpretable dimensions. Figure 1 shows the first four dimensions with selected French verb stems suitable for illustrating the dimensions. Every map contains symbols for all analytic primitives, i.e. all 360 motion events. In every map some categories of the doculect to be illustrated are highlighted by particular symbols which are given in the legend. All other situations which are not covered by any of the highlighted categories are displayed by small gray circles (not given in the legend). Due to the large number of categories it is not possible to visualize all categories in all doculects. We would prefer to use color for better visibility, but cannot unfortunately use this option in the current paper for typographic reasons. In each map additional labels, such as 'go', 'come', 'enter' etc. are inserted by hand roughly characterizing the semantic domain of that region of the map based on inspection of distribution of the language-particular categories.

The amount of data reduction in MDS analyses is illustrated in Figure 2, where the so-called *Eigenvalues* of the first 30 dimensions in the MDS analysis for verb stems are given. With 360 analytic primitives, there is a maximum of 359 possible dimensions ($n - 1$). The relative magnitudes of Eigenvalues indicate the relative contribution of the corresponding dimension in reproducing the original distance matrix. (Some Eigenvalues in high dimensions are negative, indicating that the original distances are not strictly Euclidean). Only such dimensions should be used whose Eigenvalues are larger in magnitude than the largest negative Eigenvalue. (The magnitude of the largest negative Eigenvalue is shown as a line in Figure 2.) The figure shows that at least the first 30 dimensions are of interest statistically. It also shows that Dimension 2 displays about 60% of the amount of information of Dimension 1 while Dimension 10 still displays about 15% of the amount of information of Dimension 1.

Table 4 lists the rough interpretation (using English terms) of the major semantic correlates of the first twelve dimensions. The orientation of the poles (positive vs. negative) is irrelevant, and in most dimensions there is only one pole that adds a new interpretable cluster. This is because except for Dimension 1 every higher dimension singles out one particular lexical-semantic cluster while the large majority of contexts are simply not sensitive to the semantic distinction made in that dimension. Put differently, Dimension 2 shows the extent to which an example is sensitive to transport, Dimension 3 to movement inward, Dimension 4 to movement outward, etc.

As discussed above, the dimensions with lower numbers display more information than dimensions with higher numbers (see Figure 2). The order of appearance of lexicalization patterns across these dimensions is determined (a) by the number of tokens in the database supporting a general categorization trend and (b) by the crosslinguistic recurrence of that categorization trend. The number of tokens depends both on the selection of clauses and the selection of

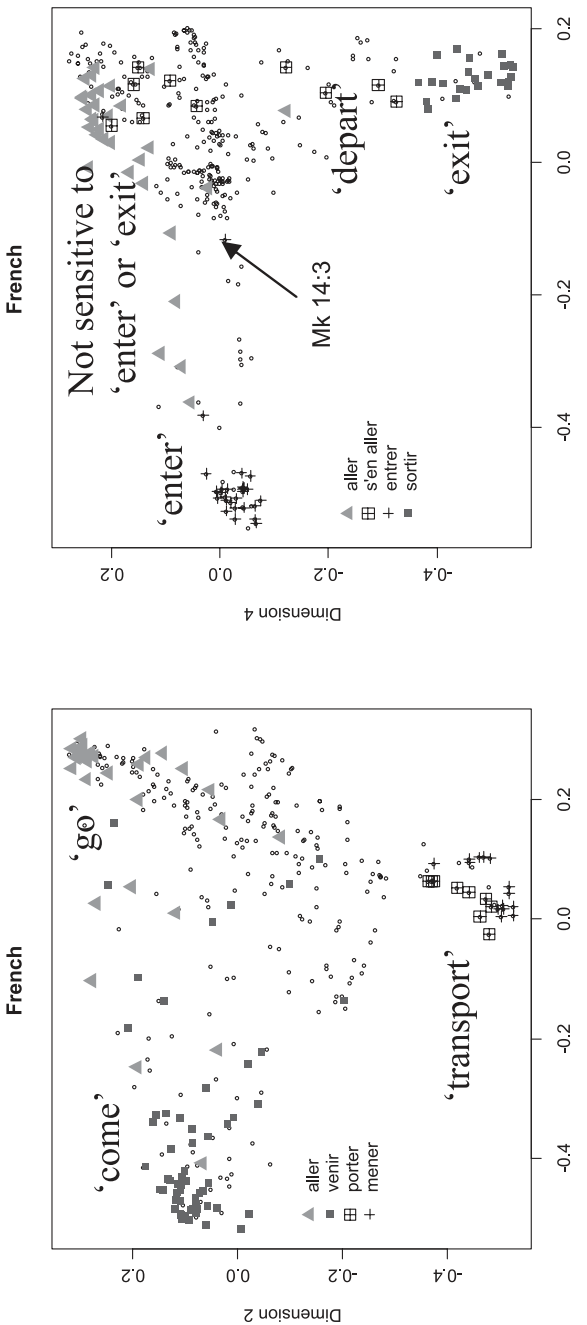

Figure 1. *The first four dimensions illustrated with French categories*

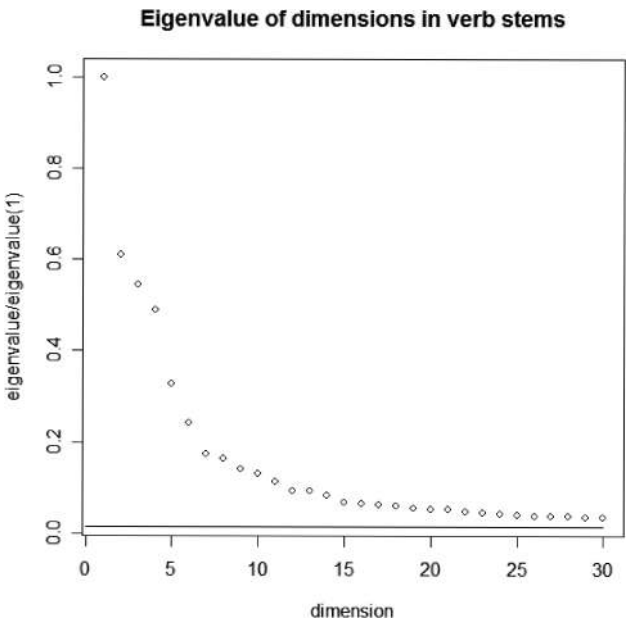

Figure 2. *The Eigenvalues of the first 30 dimensions in the MDS analysis*

Table 4. *Interpretation of the first twelve dimensions*

	Negative Pole	Positive Pole		Negative Pole	Positive Pole
1	come/arrive	go/depart	7	walk/pass/cross	
2	transport (bring)		8	assemble	
3	enter		9	ascend/descend	
4	exit/depart		10	come	arrive
5	follow		11	walk	pass/cross
6	run/flee		12	carry	lead

doculects. For example, 'enter, go in' (Dimension 3, negative pole) and 'exit, go out' (Dimension 4, negative pole) appear in lower dimensions in comparison to 'ascend, go up' (Dimension 13, negative pole) and 'descend, go down' (Dimension 13, positive pole) mainly because 'enter' and 'exit' contexts are much more frequent in Mark than 'ascend' and 'descend' contexts. 'Ascend' and 'descend' first occur together in the positive pole of Dimension 9 and are distinguished only in Dimension 13, because the sample happens to contain many European languages where there are '-scend' verb stems used both in 'ascend' and 'descend' contexts (e.g., Bernese German *styge*, Classical Greek *-baínō*, Latvian *kāpt*). However, the reason why 'assemble' appears in a lower

dimension than ‘arrive’, although there are more ‘arrive’ clauses than ‘assemble’ clauses in the sample, is because ‘assemble’-like verbs cluster more strongly crosslinguistically than ‘arrive’-like verbs.

In Table 4 the semantic correlates of the dimensions of the MDS analysis are given in the form of very general labels. It is not possible in this paper to give a detailed semantic analysis of every dimension. Dimensions 1 and 10 will be discussed in more detail below in Section 7. What is important to note here is that higher dimensions make finer distinctions. Thus, Dimension 1 is a very general distinction between ‘come’ and ‘go’ where the more strictly deictic examples are at the poles with less strictly deictic examples forming a scale between ‘come’ and ‘go’. Dimension 2 singles out transport in general (labeled here with the general English transport verb ‘bring’). A finer distinction within the domain of transport is made only in Dimension 12 in which ‘carry’ (the theme does not move by itself) is opposed to ‘lead’ (the theme moves by itself while being transported). ‘Carry’ vs. ‘lead’ is not the only possible subdivision of the domain of transport. It just happens to be better represented in the form classes in the doculects of the database than other distinctions, such as ‘transport inanimate’ vs. ‘transport animate’ (for instance, in Aymara and in Hopi, Lak, and Navajo, the latter three not represented in the sample).

Dimension 3 (negative pole) singles out inward motion. Figure 1 (right) shows that typical ‘enter’ contexts cluster densely at the negative pole of Dimension 3. However, there is one context outside of this cluster in which *entrer* is used in the French doculect, shown here in (1). In this passage, there is undoubtedly an inward movement implied, but it is so backgrounded in the context that it is rarely encoded by ‘enter’ verbs in the doculects sampled.

(1) French

. . . une	femme	entr-a	avec	un	flacon	d’	albâtre . . .
a:F	woman	enter-PV:3SG	with	a:M	box	of	alabaster

‘[And being in Bethany in the house of Simon the leper, as he sat at meat,] there *came* a woman having an alabaster box’.
(Mark 14:3)

‘Depart’ does not emerge as a dimension of its own, but rather as an intermediate area in Dimension 4 between ‘exit’ (negative pole) and ‘go’ as illustrated in Figure 1 (right) where *s’en aller* ‘depart’ ranges between *sortir* ‘exit’ and *aller* ‘go’.

The very general nature of the poles is exemplified for instance by Dimension 7 where ‘walk’, ‘pass’ and ‘cross’ verbs all cluster at the same pole. This is because the manner of going on foot (‘walk’) is often expressed by the same verb as the path through, along or across (‘pass/cross’). Dimension 11 splits this cluster into two opposite poles ‘walk’ vs. ‘pass/cross’.

It is important to note that not all lexicalization patterns found in each of the 101 texts are reflected as dimensions in the MDS. Special lexicalization patterns occurring only in one language of the sample, such as, for example, Classical Greek *poreúomai* 'go for some longer trip, travel' (see Section 7 below), never occur as poles of a dimension. Put differently, the method is good for detecting frequently recurrent lexicalization patterns, but it cannot be used to identify all lexicalization patterns in the sample.

The semantic typology of motion verbs is very complex. The MDS analysis shows that ten to twenty dimensions are needed to capture the most general, crosslinguistically recurrent lexicalizations. Other lexicalization patterns do not even emerge in the MDS analysis as dimensions because they are restricted to single languages in the sample. Moreover, the text considered does not reflect the full diversity of the motion verb domain. Many contexts simply do not occur in the text (for example 'go by vehicle' is only represented as 'go by boat', there is no 'going upstream' or 'downstream', there is no 'climbing trees'). Other aspects of motion events are weakly represented, because translation does not favor them. This holds especially for motion verbs used in absolute frames of reference (Levinson 2003).[18]

It is important to note that the number of interpretable dimensions obtained does not derive from any MDS-settings but is data driven. Since there are few similar investigations it is difficult to assess how motion events relate to other domains of lexical typology in terms of number of dimensions needed in MDS analyses. However, in grammatical domains, such as local phrase markers (Wälchli 2010) or tense-aspect categories (Croft and Poole 2008), two dimensions are usually sufficient. We can derive from this a hypothesis that lexical domains tend to be more multidimensional than grammatical ones. However, to verify this hypothesis many more domains will have to be analyzed with probabilistic semantic maps.

7. More detail: a semantic map of 'go', 'come' and 'arrive'

We will now argue on the basis of 'go', 'come' and 'arrive' verbs that lexical typology indeed needs massively crosslinguistic approaches because the amount of typological diversity is very high, while at the same time there are strong regularities in the form of major trends. At least some major trends, we will argue, have obvious semantic correlates, but are also shaped by discourse. Further, recurrent aspects of diachronic developments can be identified, even though diachronically there is also a high amount of diversity. Finally, most categories can be attributed to recurrent category types (even though they are crosslinguistically only similar but not identical), but there are also many rare category types attested in only a single or a few languages. All this will be

shown on the basis of the MDS analysis of data from parallel texts, which is why we think this approach to lexical semantics is highly useful both for theory and description.

Our focus in this section is not to give a full account of 'go', 'come' and 'arrive' verbs, but the following exposition is more of an exemplary nature to illustrate the general points that we want to make in this paper. We could equally well have picked others aspects of motion encoding. For the same reason, this section does not provide a comprehensive survey on earlier typological research in deictic motion verbs, but only make incidental references to the large amount of work in this extensive domain of scholarly debate.

In the MDS, Dimension 1 roughly distinguishes 'come' from 'go', and contexts describing 'arrive' are distinguished in Dimension 10 (see Table 4 above). For this section, we selected these two dimensions as x- and y-axes respectively for our visualization, which gives us a two-dimensional constellation of all 360 situations to be compared across the languages of the sample. Figure 3 illustrates this semantic map for four doculects. The x-axis shows Dimension 1, which distinguishes 'come' contexts (negative pole on the left) from 'go' contexts (positive pole on the right) and the y-axis shows Dimension 10 which distinguishes 'arrive' contexts at the positive pole on top. For convenience, typical 'come'-like lexemes are given as dark squares, typical 'go'-like lexemes as gray triangles, and typical 'arrive'-like lexemes as window symbols. In every map there is a legend added.

The three labels 'go', 'come' and 'arrive' in the maps stand for lexical domains that are located in that region of the map. We use the label 'come' in the bottom left corner because lexemes are predominantly found here that are all given the meaning 'come' in dictionaries and grammars: Acholi *biínô*, Albanian *vij*, Classical Armenian *gam*, Aymara *juta*, Bambara *na*, Basque *etorri*, Cakchiquel *pe*, English *come*, Estonian *tulema*, etc. In the same way, the top is labeled 'arrive', because here we find Albanian *arríj*, Aymara *puri*, Bambara *se*, Basque *iritsi*, etc., which all mean 'arrive'. It might be objected that we are just taking the labels from dictionaries and grammars and that the semantic interpretation hence does not derive from the statistical analysis. However, note that we do not label categories in individual languages, but we label a region of the semantic space that is the same for all doculects of the sample. Thus, the English map has an 'arrive' corner like all other maps even though the verb *arrive* is not attested in the whole text considered. We could have used labels that are entirely arbitrary, like "A77" or "164735". The reason why we label the corners of the triangle 'go', 'come' and 'arrive' is because this makes the analysis compatible with existing linguistic work in most different languages in the same way as it is convenient to label vowels in formant diagrams in acoustic phonetics with IPA vowel symbols. Most importantly, the method allows us to plot all the different 'come' categories, as there is no pair

of doculects where 'come' verbs have exactly the same extension. There are 101 maps, one for every doculect. However, we can discuss only a few of them here.

It is not trivial to clarify the distinction between 'arrive' and 'come' descriptively. Quantitatively the MDS plots show that this distinction is less clear-cut than various other distinctions in motion events. What distinguishes contexts where 'arrive' verbs occur most frequently in many doculects are especially two aspects of perspective which can apply individually or jointly: (a) 'arrive' is used for moving to places which have been previously established in discourse rather than for introducing new places, and (b) arrival is a nontrivial achievement (i.e. it is beyond the full control of the figure whether or when s/he will arrive). For instance, Mark 5:1 *And they came over unto the other side of the sea* is a typical 'arrive' context because the new place has been introduced already before in 4:35 (*Let us pass over unto the other side*) and there has in fact been an unexpected storm on the sea which is why it was not clear for everybody whether they actually could make it to the other side.

Every lexeme in each of the four doculects in Figure 3 has its individual categorization pattern. This is reflected in the maps by the different distributions of lexemes. For every map the squares and triangles exhibit a specific individual distribution. Crucial is that the configuration of situations is kept constant across all languages. This could also be done by arranging the situations in some other way. For example, one could use their linear sequence of occurrence in the Mark text as shown in Figure 4. Figure 4 shows exactly the same categories in the same situations as Figure 3 in a crosslinguistic constant distribution. The difference is that unlike Figure 3 the situations are not arranged according to their semantic similarity, which makes it very difficult to see how the categories cluster. Using the dimensions of the MDS analysis has the clear advantage that pairs of close dots are more likely to be expressed by the same category in any language. Because of this, the visual impression of the MDS display is much more informative than the linear order of Figure 4.

Comparing the maps in Figure 3, there are various observations to be made about differences and similarities between the four languages shown. For example, when looking at the y-axis, *King James* English does not have any opposition between a 'come'-verb and an 'arrive'-verb: there is only 'come' (square symbols) all over in the top and bottom left corners. In *Reina Valera Antigua* Spanish the situation is similar, but there are three *llegar* dots interspersed in the field of *venir* (square symbols) that cover both the 'come' and the 'arrive' domains. In Hungarian and *Lenguaje Sencillo* Spanish, however, there is a clear opposition between 'come' and 'arrive' since the two domains are covered by different category symbols (which have, however,

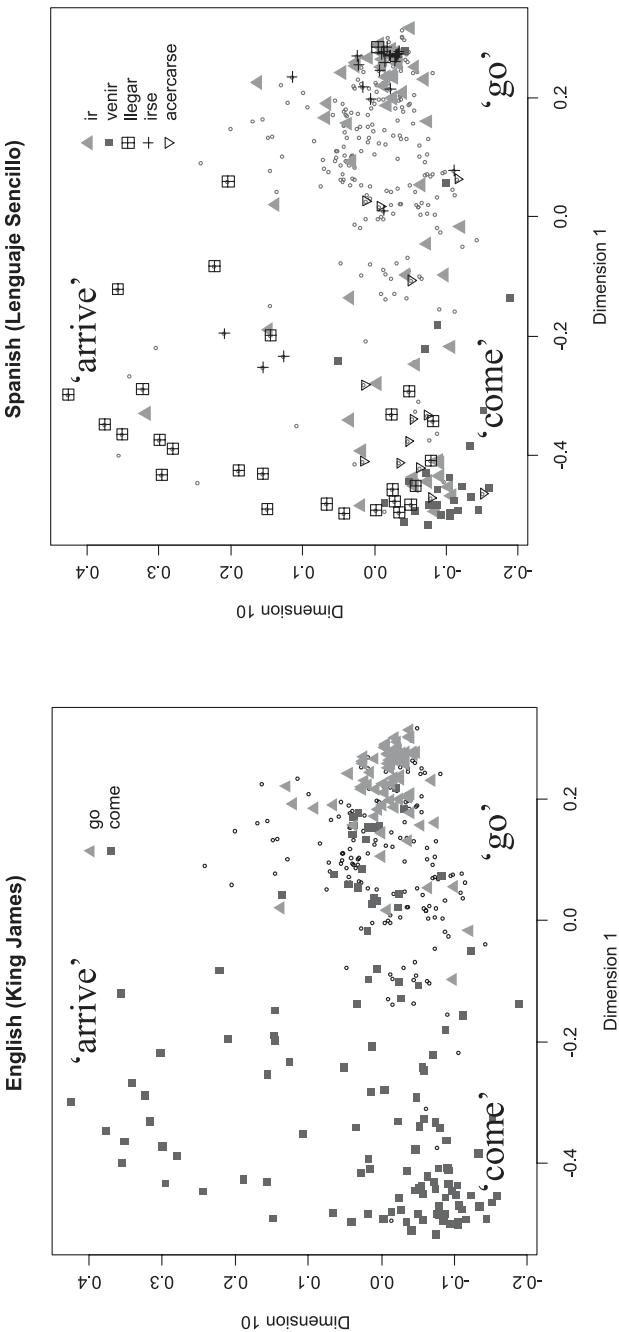

Figure 3. *A semantic map of 'go', 'come', and 'arrive' in "fully deictic" (Modern Spanish and Hungarian) and "predominantly-deictic" languages (English and 16th century Spanish).*

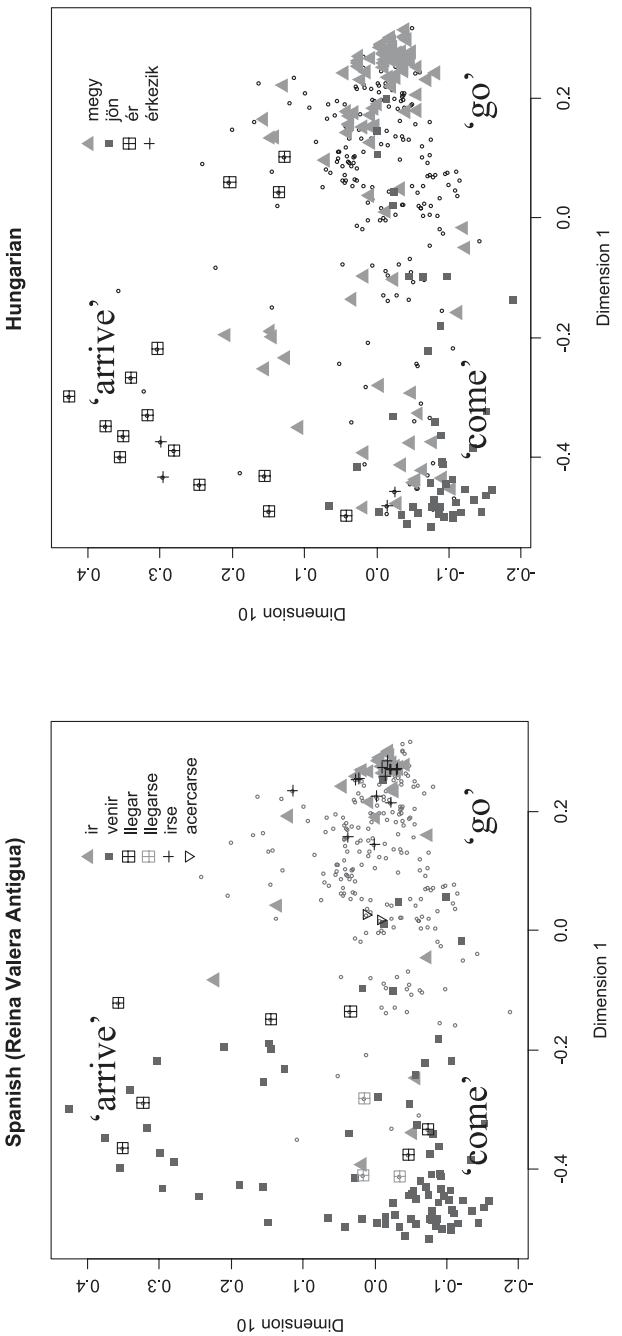

Figure 3. *(Continued)*

Figure 4. *The 360 motion event situations arranged in their linear order in the text (from top to bottom, then left to right).*
*Some readers find it difficult to understand why dots arranged in a square can be linear. They are linear like the lines on a book page are linear (but read from top to bottom first). Plotting all symbols on one line would require a 2m foldout paper, which is, of course, not practical. The symbols are the same as in Figure 3.

slightly different extensions in Hungarian and *Lenguaje Sencillo* Spanish). When focusing on the 'go'-like contexts (in the bottom right of the maps), *King James* English and Hungarian both show one single predominant 'go'-verb. Both variants of Spanish distinguish between the verb *ir* 'go' and the reflexive *irse* 'depart' in this domain. But, for showing the opposition between 'go' and 'depart' we would need another dimension (Dimension 4, see Figure 1, right). Further, at the bottom left of the map, both Hungarian and *Lenguaje Sencillo* Spanish show a variety of lexemes being used in the 'come'-like contexts, in contrast to *King James* English and *Reina Valera Antigua* Spanish, in which almost all contexts in the bottom left are coded by the same lexeme.

As pointed out by Kaiser (2002, 2005) and Dahl (2007), Bible translations have the advantage that they allow us to trace diachronic developments, if translations from various stages are attested. For example, the restricted usage of Modern Spanish *venir* (only used in strictly deictic 'come' contexts) is a recent development, as previously observed by Ricca (1993: 131) and de Icaza (1916). This development can clearly be shown on the basis of translations of Mark. As a representation of Modern Spanish, the Bible translation *en lenguaje sencillo* has been selected because it is little influenced by the 16th century *Reina Valera Antigua* translation. Figure 3 shows that 16th century Spanish *venir* had a much wider usage than in Modern Spanish. The diversity of the usage of *venir* in 16th century Spanish looks much like *come* in Early Modern English (*King James* translation), and contrasts with *venir* in Modern Spanish (see Table 5 for examples). English did not undergo a similar shift, though the use of *come* in *King James* English is slightly different from its use in Modern English (not shown here). The usage of *come* in *King James* English is more akin to that of German *kommen*, Swedish *komma* and Icelandic *koma* than to the usage of *come* in Modern English. These results are in accordance with Ricca (1993) where the differences in deictic motion verbs in European languages are discussed in great detail.

Given the observed diversity in use it is difficult to understand how we can identify English *come* and Spanish *venir* and say that they are simply instances of a category type of 'come' verbs. This situation is representative of any pair of doculects in the sample in the sense that it is mostly extremely difficult to match categories by distribution manually if we do not somehow assume that the two categories "mean" the same thing. The constellation in Figure 3 suggests that the difference between 'go' and 'come' can be viewed as a scale: 'come' and 'go' verbs generally cluster at the poles, but languages differ in how the transition area is encoded. In English this area is covered mainly by the category *come*, in Modern *Lenguaje Sencillo* Spanish mainly by *ir* 'go'. For example, Table 5 singles out seven contexts where English uses *come* while Modern Spanish uses *ir*. Each of these contexts has a particular meaning

of its own, which cannot be fully expressed by the short excerpt of the examples given in the table, without quoting the full passage in which it is embedded. However, all contexts have in common that they express motion to the next place of the story line where the narrated action continues. We call this type of context *narrative 'come'*.

Narrative 'come' is neither centripetal nor centrifugal and therefore not strictly deictic and has thus not been in the focus of studies of deixis in motion verbs. It is clearly related to Ricca's (1993) predominantly deictic 'come' in that it is goal oriented, but it is not centripetal in the same way as motion toward the speaker.[19]

Table 5. *The "pseudo-deictic" domain exemplified by Spanish* ir *and English* come.

	English (*King James*)	Spanish (*Biblia en lenguaje sencillo*)
1:14	*Jesus* ***came*** *into Galilee*	*Jesús* ***fue*** *a la región de Galilea.*
1:31	*And he* ***came*** *and took her by the hand . . .*	*Jesús* ***fue*** *a verla . . .*
6:29	*They* ***came*** *and took up his corpse . . .*	***fueron*** *a recoger el cuerpo de Juan . . .*
6:48	*he* ***cometh*** *unto them, walking upon the sea.*	*Jesús* ***fue*** *hacia ellos caminando sobre el agua*
12:14	*And when they were* ***come*** *. . .*	*Ellos* ***fueron*** *y le dijeron:*
14:17	*And in the evening he* ***cometh*** *with the twelve.*	*Jesús y los doce discípulos* ***fueron*** *al salón.*
16:2	*they* ***came*** *unto the sepulchre . . .*	***fueron*** *a la tumba de Jesús.*

There are also languages where the narrative 'come' domain is represented by a separate third verb. Examples are the languages Mari (Uralic) and Chuvash (Turkic), as illustrated in Figure 5. The discussion of these two languages is interesting both diachronically and areally. In both Mari and Chuvash the intermediate 'narrative come' verb is an erstwhile 'go' verb. Mari *mijaš* is related to Finnish *mennä* 'go' and Hungarian *megy* 'go' (Rédei 1988: 272), while Chuvash *pyr-* is related to Yakut *bar-* 'go'.[20] Though from different genealogical origin, Mari and Chuvash are in close contact. There is direct evidence for contact-induced language change as Chuvash *kaj-* 'go' must be related to Mari *kajaš*. The direction of the borrowing is not entirely clear.[21] However, the semantic maps here help to interpret the diachronic development. We hypothesize that the Mari verb *kajaš* (originally meaning more specifically 'walk') enters the 'go' domain and pushes the erstwhile 'go' verb *mijaš* toward 'come' to become a pseudo-deictic verb. A similar development must have occurred in Chuvash with *pyr-* (originally 'go') being pushed toward 'come' by *kaj-*. This looks very much like a push chain known from historical phonology (see Labov 1994). This development is summarized in a more schematic form in Figure 6.

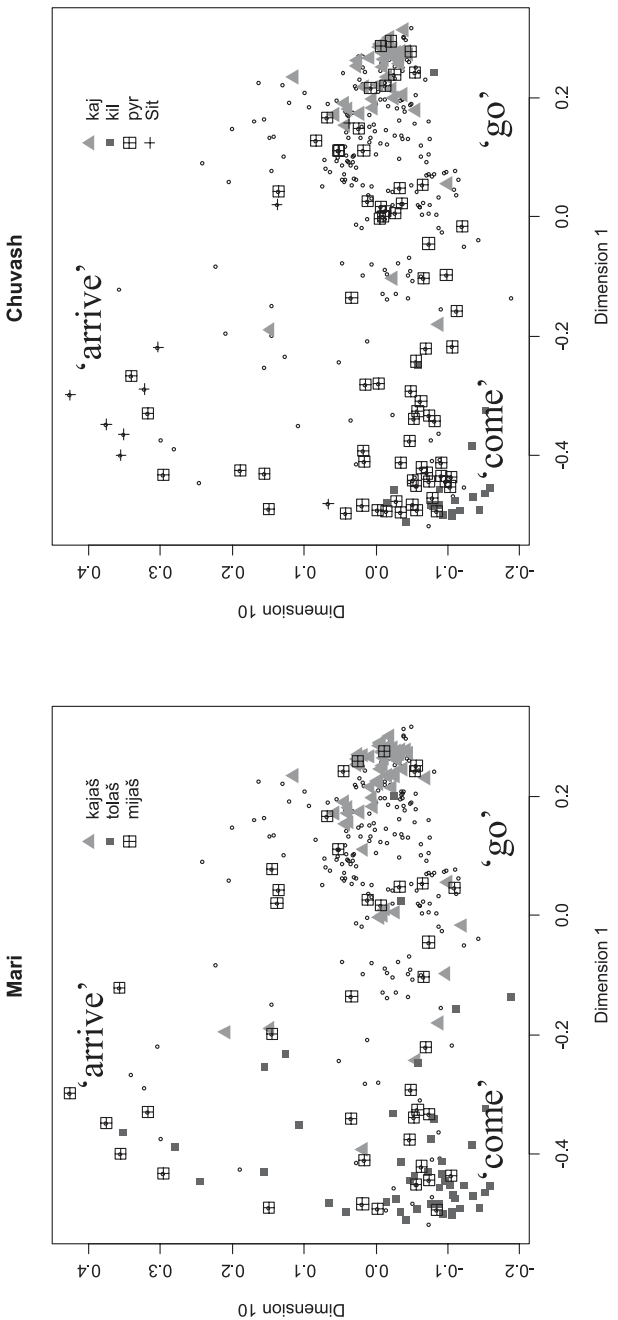

Figure 5. *‘Go’, ‘come’ and ‘arrive’ verbs in Mari and Chuvash*

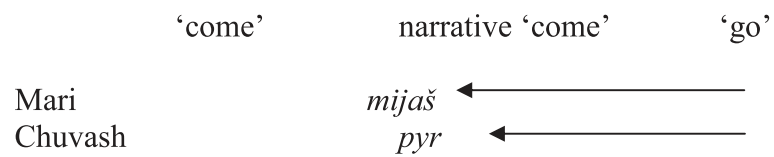

Figure 6. *Parallel semantic shift of Mari and Chuvash erstwhile ‘go’-verbs*

Narrative ‘come’ is not the only attested category type beyond the cardinal category types ‘go’, ‘come’ and ‘arrive’. Some American languages have two ‘arrive’ verbs which are often conveniently glossed as ‘arrive here’ and ‘arrive there’. An example is San Miguel el Grande Mixtec, given in Figure 7 (left). The semantic map indeed supports the idea that *caà* ‘arrive here’ [given in the legend of the figure as *Caa*] is more centripetal than *xaà* ‘arrive there’ [given in the legend of the figure as *xaa*]. The two verbs differ in their arrangement on Dimension 1 (x-axis) at least as a tendency. However, the two verbs cannot be considered in isolation; equally important is the opposition between *caà* ‘arrive here’ and *kii* ‘come’. The latter is a very restricted properly deictic ‘come’ verb. While *kii* is used in direct speech (motion toward speaker), *caà* ‘arrive here’ is used in narrative contexts (motion toward the deictic focus of the story). This contrast is shown in (2) in which (2a) with *kii* ‘come’ is from direct speech and (2b) with *caà* ‘arrive here’ is from a narrative sequence.

(2) Mixtec (San Miguel el Grande) (Oto-Manguean, Mixtecan)

a. . . . éliá, a nī **kii**-de
Elias, ? COMPL come-3.M
‘. . . Elias is indeed come . . .’
(Mark 9:13)

b. . . . te nī **chaā** tΛ-saā . . .
and COMPL arrive.here CL-bird
‘the fowls of the air came . . .’
(Mark 4:4)

Mapudungun (Araucanian; Chile) has yet another system: *küpa-* is strictly deictic ‘come’, *puw-* is ‘arrive’. Further there is a space for a third intermediate verb *aku-* between them, which is not fully adequately glossed with ‘arrive here’. While it is not possible to present all the different categorization patterns found in the doculects sampled, it is important to note that there are many more ‘arrive’ verb patterns that are all slightly different.

The examples given have served to illustrate the great diversity in lexical semantics of motion verbs across different languages. Finally, let us look at two languages where categorical distinctions of motion verbs are not supported by the perspective obtained by the MDS analysis because the kinds of

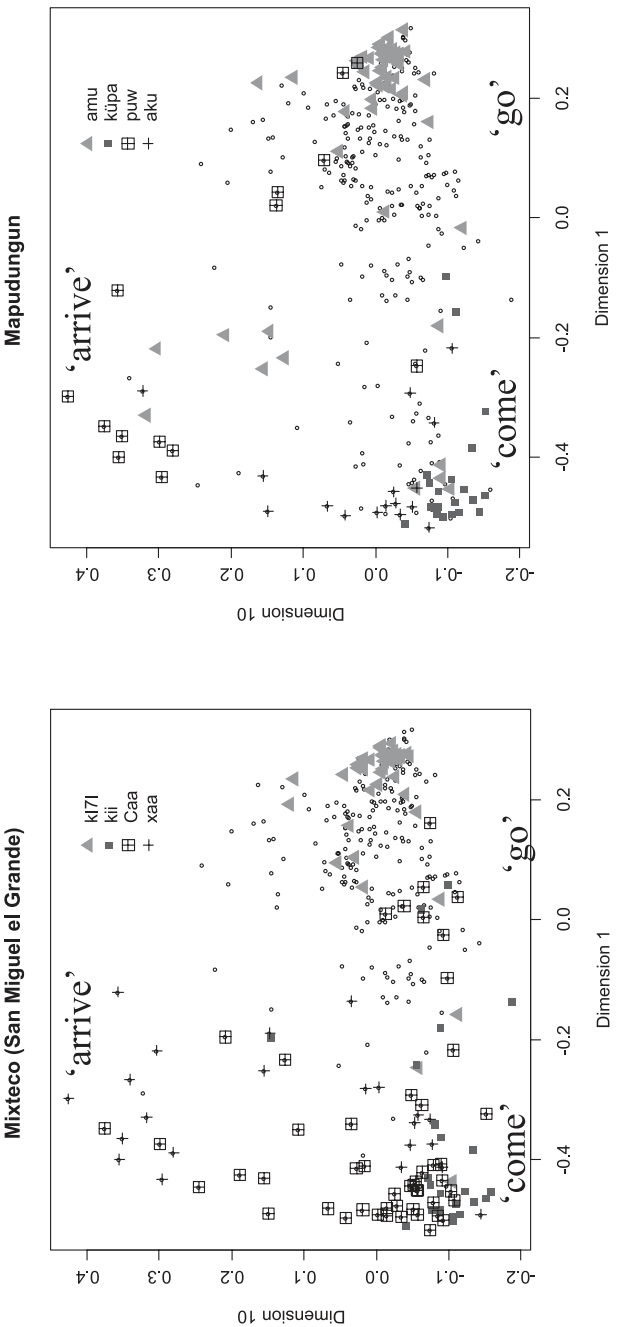

Figure 7. *'Go', 'come' and 'arrive' verbs in Mixtec and Mapudungun*

categorization made are not recurrent in the sample (cf. Figure 8). First, the Classical Greek *poreúomai* approximately means 'travel', though this is not an adequate translation, rather it expresses a motion that can somehow qualify as a trip, not just simply going somewhere. As can be seen in the figure, the verb is used in a small part of the domain of the 'go' contexts, while the verb *érxomai* simply encompasses the whole 'go/come/arrive' domain. The specific distribution of *poreúomai* does not have any close correlate in any other language of the sample, even though the Classical Greek text is the direct or indirect source of all translations used here.

Second, in Sora (Austroasiatic, Munda), there are two verbs *iy-* and *yer-*, both translated as 'go' in Ramamurti (1986). The additional 'hither'-deixis is expressed by a suffix that can be combined with both stems which is not considered here. We cannot tell what the semantic difference is between the two verbs. However, what we can say for sure is that this type of categorical distinction is not made in any other language of the sample in this way. It is thus one of many rare category distinctions attested in the sample. Figure 8 also allows us to determine that Sora has a relatively straightforward 'arrive' verb *ardu-*.

Let us now summarize this section according to the list of requirements given in Section 4. The semantic map is massively crosslinguistic in the sense that the configuration of situations is based on a large sample of doculects. However, at the same time it is language specific in that it can map the form classes of each individual doculect in the sample. The amount of diversity to be accounted for is high; there are several form classes in every doculect and there are very few pairs of form classes in the database with identical distributions. However, at the same time there is a high amount of regularity. Most 'go/come/arrive'-verbs can be neatly located in some area of a triangle with 'go', 'come' and 'arrive' as cardinal points. Many of the different category types can be shown to have semantic correlates. The narrative discourse structure of the underlying text plays a major role. It highlights a particular complex of situations termed here narrative 'come' which is usually neglected in studies of deictic motion verbs. The semantic map shows that narrative 'come' is a diffuse transition zone between cardinal points rather than a clear-cut cluster. In most languages it has no exclusive form class but is rather expressed by a 'come' verb or a 'go' verb. However, under particular diachronic constellations, when an erstwhile 'go' verb is pushed toward 'come' — as happened in the contact languages Mari and Chuvash — it can be expressed by a form class of its own. Finally, the method presented is no strict quantitative tool for the partitioning of all form classes into categorization types, but rather has a heuristic function for identifying possible category types. We start with identifying typical category types in the cardinal points of MDS-plots and then further proceed to detect more specific category types, such as narrative 'come' in

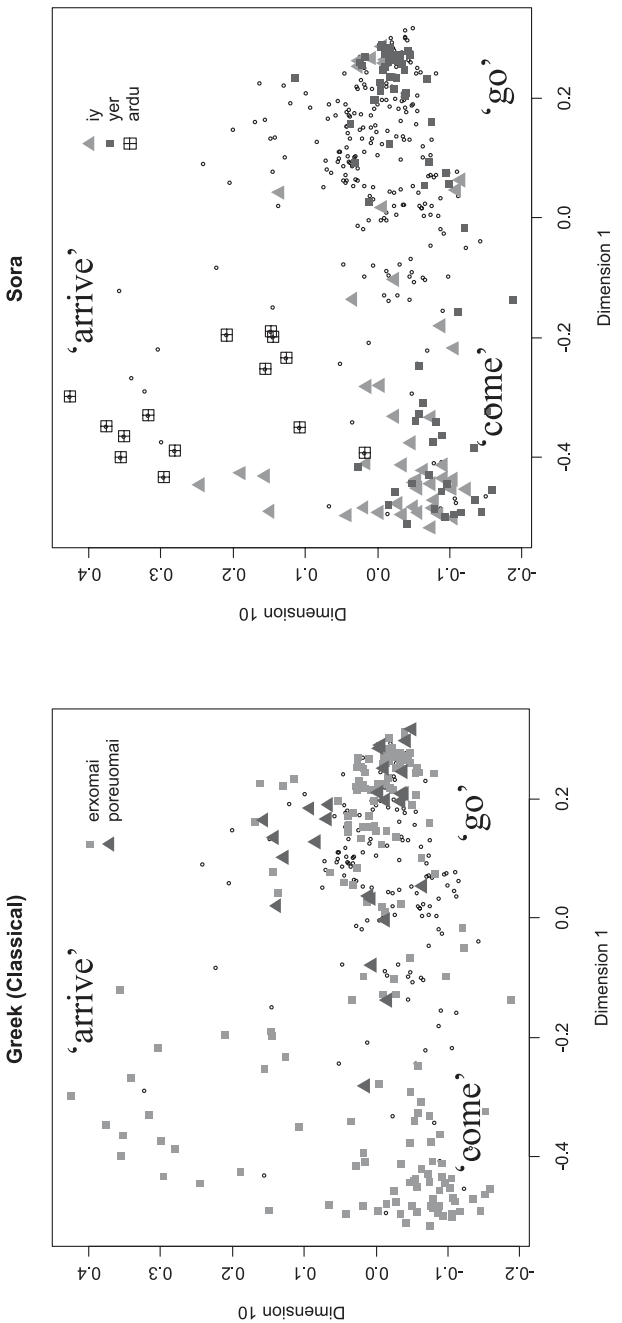

Figure 8. *The 'go', 'come' and 'arrive' domain in Classical Greek and Sora*

Mari and Chuvash, two kinds of 'arrive' verbs in Mixtec, and an intermediate verb between 'arrive' and 'come' in Mapudungun.

9. Conclusion: toward a typology without types

One of the main advantages of the traditional semantic map approach is that no decision has to be made between a polysemist and a monosemist position (Haspelmath 2003: 213). The approach presented here goes a step further. Not even polysemy and homonymy need be strictly distinguished. In implicational semantic maps, when cases of homonymous categories (i.e., semantically completely different categories) are interpreted as identical this may lead to a complete breakdown of the method because rarely attested connections are assigned much weight (see Cysouw 2007). In contrast, MDS plots can visualize both macrotypological distinctions and microtypological variation at the same time. They are thus a finer tool for crosslinguistic comparison than a classification into rough types. They also show that crosslinguistic comparison is possible without abstraction of types at any level of analysis. Semantic maps from exemplar data enable us to practice a typology without types.

At the cost of a loss of a certain amount of idiomaticity and perhaps even some systematic losses due to effects of translation, constructing semantic maps from exemplar data in parallel texts has several advantages: (a) it implements in concrete terms the functional typologist's abstract ideal of translation equivalence of functional domains; and (b) it allows for a higher resolution of analytical primitives than in the case of precategorized functions based on data from reference grammars. The mapping method can be easily implemented by standard techniques of statistical analysis (computing distance matrices with a distance measure such as the Hamming distance, visualization tools such as classical multidimensional scaling). However, it has to be emphasized that the mapping relations are never one-to-one. There are always different possible ways of analysis, each with its particular advantages and disadvantages. Semantic maps will thus never reflect *the* semantic space, if there is such a thing at all. Yet, the method exemplified here provides a convenient tool for the crosslinguistic description and analysis of semantic differences of categories in different languages.

The method used here is equally applicable to grammatical as well as to lexical domains. However, it is especially relevant for lexical typology, because the lexicon is more difficult to investigate on the basis of reference grammars and traditional dictionaries.

As in many other sciences, in typology there is a tension between those who want to reduce everything to a few grand laws and those who are more interested in doing justice to individual facts. The semantic map approach has the

great advantage of having the potential to reconcile the two opposite aims which are both equally noble. The great rough picture emerges for those who are interested in general principles, but the details can be retained at the same time. Caring about details is not "butterfly collection". Keeping as much detail information as long as possible — even throughout advanced stages of analysis — is crucial because we never know if what we believe to be the relevant features really are the only essential ones. The approach proposed in this paper opens the way for a typology where generalizations can be made without there being any need to reduce the attested diversity of categorization patterns to discrete types.

Received 24 April 2010
Revised version received 18 March 2011

Stockholm University
Ludwig Maximilians University Munich

Appendix A. Semantic maps and diachrony

It has been argued that "the best synchronic semantic map is a diachronic one" and "the best semantic map is a *semantic* semantic map" (van der Auwera 2008: 43). First of all, we would like to point out that we agree with van der Auwera (2008: 45, 39) that we need both diachronic and synchronic maps and that the MDS method is unlikely to take over and to replace classical semantic maps. The reason for the latter point is simply that the MDS method needs large databases of exemplar data from many languages to become really powerful and such databases are not widely available.

We do not, however, share the point of view that diachronic semantic maps are a completely different kind of semantic maps from synchronic maps. Diachronic maps are always also synchronic maps, since they use synchronic multiple use patterns as an indirect method to reconstruct diachrony. The diachronic maps of van der Auwera and Plungian (1998) rest on (grammaticalization) paths which they connect and extend (van der Auwera and Plungian 1998: 87). The direct diachronic evidence for paths is highly restricted. Hence grammaticalization studies mainly use indirect synchronic evidence for diachrony: "multiple uses . . . can be employed as diagnostics of earlier history . . ." (Bybee et al. 1994: 18), "synchronic multiple meanings of a single marker must be seen as stages on a path" (van der Auwera and Plungian 1998: 111). Hence, so called diachronic evidence is often synchronic evidence. This point is also important for the homonymy discussion to which we turn now.

Van der Auwera (2008: 41) argues that diachronic maps are semantically optimal because of homonymy. For van der Auwera and Plungian's (1998), polysemy is direct connection of adjacent meanings whereas homonymy is formal identity without connectedness in diachronic semantic maps. Diachronic maps, it is argued, reflect the difference between homonymy and polysemy more accurately than maps relying on synchronic multiple use patterns only because "relaxing adjacency owing to the disappearance of a link meaning" (van der Auwera and Plungian 1998: 113) can disturb the

picture. The typical example adduced is the minimap HABITUAL → PROGRESSIVE → FUTURE where the present progressive is sometimes replaced by another form (as in Turkish *okut-ur* 'uses to teach/will teach' vs. *okut-uyor* 'is teaching'; Haspelmath 2003: 236; van der Auwera and Plungian 1998: 113). First of all it must be said that the exclusivity of this path is not established beyond any doubt, since there is at least one case of a future-habitual polysemy without any progressive present involved. The Žemaitian Lithuanian habitual past is formed with the future tense (Wälchli 2011). Like all diachronic reconstructions, diachronic semantic maps are highly speculative. Hence, if we want to define homonymy and polysemy in purely diachronic terms we are highly dependent on synchronic data. Usually typologists use crosslinguistic evidence to distinguish homonymy and polysemy: "if many diverse languages independently have the same pattern of 'homonymy', then the meanings are closely related" (Croft 2003: 106). This is Haiman's *Isomorphism Principle*. There is no semantic map method of whatever kind that does not implicitly or explicitly rely on the idea that formal identity reflects semantic similarity because of iconicity. Neither accidental multiple use patterns (homonymy) nor nonaccidental ones (polysemy) are rare, but polysemy patterns are iconic (exhibit parallels between form and meaning), and since meaning is largely universal, polysemy patterns tend to follow the same paths in most different languages over and over again while homonymy patterns are accidental.

The major difference between the traditional and the MDS approach is the treatment of exceptions (the nonrecurrent patterns). The traditionalists think the exceptions should be excluded before the maps are drawn. We think we should use a method that reflects the major trends in the data where the recurrent tendencies are much more strongly reflected than the accidental patterns. Our advantage is that we need not distinguish homonymy from polysemy — which is often impossible. For example, in a diachronic path A → B → C, the AC polysemy pattern will be rare while the AB, BC and ABC patterns will be frequent. As a result, the MDS will still plot these categories on a line A-B-C.

A reviewer argues: "If homonymous *blabla* 'potato' and *blabla* 'toe' are both allowed on a semantic map then this map is no longer meaning based, but form based. Of course, the method will show that in most languages no one form is used for both 'potato' and 'toe'. So the heuristic value is by no means denied. Yet homonymy-allowing maps are thus not semantics." Since all semantic maps rely on identity of form (sometimes very implicitly, it is true), we do not agree that diachronic maps are *semantic* semantic maps in the sense of truly semantic maps. Semantic maps are always an indirect approach to semantics by means of form (multiple use patterns). There are hence no *semantic* semantic maps. Semantic maps is semantics as dendrochronology is chronology: indirectly. What we measure is formal identity or annual rings in wood and what we aim at is meaning or time. This is possible because meaning and form are related in a particular way that is stated explicitly (the isomorphism hypothesis) in the same way as dendrochronologists argue that annual rings and time are related in particular ways. The utility of the method depends on whether the relation of what is measured and what is aimed at is dominant over anomalies that disturb the relation. Anomalies — such as homonymy discussed above or the different behavior of zero marking (Malchukov 2010; Wälchli 2010) — can be addressed in various ways by different researchers. This is the

kind of discussion we must advance to improve the method. But it is not the case that any one approach of those discussed so far is fundamentally different from all the other ones in being much more "semantic".

Finally, it is not clear to us whether "diachronic maps" with their focus on extension of polysemy patterns reflect the full range of diachronic phenomena related to semantic change. Polysemy patterns can also be lost and this may be relevant for semantics. If grammaticalization leads to polysemy extending over a large area of conceptual space it is not iconic any more to maintain formal identity between the most distant chain links. Krug (2001) shows that younger age groups in the British National Corpus have a significantly higher proportion of NPs with *going to* along with a much higher incidence of contracted *gonna*, which suggests that they have further progressed in the functional split between modal (or futural) *gonna* + infinitive and spatial *going to* +NP. This reestablishes isomorphism (i.e., rather distant meanings which happen to be connected by grammaticalization become different in form again). Parallel to the split there is a formal convergence between *wanna* (< *want to*), *gonna* (< *is/am/are going to*) and *gotta* (< *have/has got to*) which have paradigmatically similar meanings ("emerging modals") despite their very different formal and semantic origin. There are many examples of this kind. *One* and *a* are not formally identical anymore because their meaning — even though related — is quite different. The same holds for French *avoir* 'have' and the endings of the future tense.

Hence, if we say that pairs of closer dots are more likely to be expressed by the same meaning, this has two diachronic counterparts: (i) Different forms in a pair of meanings are more likely to change into the same form the closer the meanings are related and (ii) any identical formal expression of a pair of meanings is more likely to be separated into different form the more distantly the two meanings are related.

We would like to add that several ways in which our probabilistic maps can be applied to diachronic questions have been addressed in the discussion of two stages of Spanish and of the pushing chain 'go' → narrative 'come' → 'come' in Mari and Chuvash (see Section 7). A major difference between a domain such as motion verbs and a domain such as modality is that recurrent diachronic pathways are omnipresent in the latter. We think that any lexical and grammatical domain can be studied with the semantic map method and that no precedence should be given to such domains as modality which are paved with well-known paths of grammaticalization. Semantic map research, as we understand it, has a strong heuristic component. It can help us to find yet unnoticed generalizations. This makes this method important for lexical typology which is not yet as equally well studied as grammatical typology.

Appendix B. Languages sampled

The languages are: Acholi, Ainu, Albanian, Classical Armenian, Avar, Aymara, Bambara, Basque, Cakchiquel, Chamorro, Chiquitano, Choctaw, Chuvash, Dakota, Dinka, Drehu, Efik, English (*King James*), Setu Estonian, Ewe, Fijian, Finnish, Stadin Slangi Finnish, French, Garo, Bernese Swiss German, Classical Greek, Modern Greek, Guaraní, Haitian Creole, Hausa, Hawaiian, Hindi, Hmong Njua, Hungarian, Icelandic, Ijo,

Indonesian, Irish, Italian, Jamaican Patois, Jul'hoan, Kala Lagaw Ya, Kalderash Romani, Kâte, Khalkha Mongolian, Khasi, Khoekhoe (Nama), Komi, Kuna, Kunama, Latgalian Latvian, Latin, Latvian, Lithuanian, Livonian, Maori, Mapudungun, Mari, Marshallese, Miskito, San Miguel el Grande Mixtec, Mizo, Maltese, Erzya Mordvin, Car Nicobarese, Ossetic, Papiamentu, Piro, Pitjantjatjara, Sutsilvan Rhaeto-Romance, Romanian, Russian, Saami, Samoan, Sango, Shipibo, Songhay (Koyra Chiini), Somali, Sora, Spanish (*Reina Valera Antigua*), Spanish (*Lenguaje Sencillo*), Sranan, Swahili, Swedish, Tajik, Tagalog, Toaripi, Toba Batak, Tok Pisin, Tongan, Turkish, Udmurt, Ulawa, Veps, Vietnamese, Wolof, Yabêm, Yoruba, Isthmus Zapotec, and Zulu.

Notes

* Correspondence address: Bernhard Wälchli, Stockholm University, Department of Linguistics, SE – 10691 Stockholm, Sweden. E-mail: bernhard@ling.su.se.

1. We are grateful to Maria Koptjevskaja-Tamm, Martine Vanhove, Östen Dahl, Iwar Werlen, and six anonymous reviewers for valuable comments and to Deborah Edwards for proofreading. While collecting the data, BW was supported by the Swiss National Science Foundation (PA001-104983), and while writing this paper by the German Science Foundation (DFG) in the SFB 471 "Variation and Evolution in the Lexicon" in Konstanz and the Swiss National Science Foundation (PP001-114840). Abbreviations in glosses: CL classifier, COMPL completive, F feminine, M masculine, PV past perfective, SG singular.
2. For a discussion of primary and secondary data see Lehmann (2004). The term *doculect* (documented lect) was coined by Michael Cysouw, Jeff Good, and Martin Haspelmath in 2006 at the Max Planck Institute for Evolutionary Anthropology and is first mentioned in the published literature in Bowern (2008: 8). The relation between doculect and language can be compared to the relation between sample and population in statistics. A doculect can be more or less representative of a language, and because it is not obvious how to best sample a language, we prefer to explicitly compare empirical samples of languages, rather than to assume that any particular sample fully represents a language. For example, the doculects considered in Sections 6 and 7 are translations of Mark, which often represent a somewhat special variant of the language. Whenever the term "language" is used below, it refers to a language variety as represented by this particular doculect.
3. Of course, situations are always contextually embedded. So, although the term "contextually embedded situation" is to some extent pleonastic, we think that the attribute "contextually embedded" is useful to emphasize that we do not work with isolated examples. Note, for example, that situations can be isolated (as sometimes occurs in questionnaires) and more or less connected to other situations.
4. Precluding a full discussion, which falls outside of the scope of this paper, we propose that to establish an abstract functional domain it is necessary to show that domain-internal diversity is smaller than cross-domain diversity. The idea behind this proposal is, roughly spoken, that an abstract function can be conceived of as a set of exemplars expressing similar senses. However, such a set of exemplars only constitutes a suitable crosslinguistic function when all exemplars are expressed identically in all languages. In practice, lexicalization is too variable across languages for this to happen. Changing the categorical definition into a probabilistic formulation results in a constraint on useful functions that they should be internally consistent, but maximally differentiated among each other. In practice, such a clustering can be achieved by computational approaches in the tradition of *k*-means (see, e.g., Kaufman and Rousseeuw 2005).

5. In principle, all three types of analytic primitives (i.e., doculects, form classes, situations) can be investigated separately, and biases in the samples for each of them can be detected (and corrected for) post-hoc by resampling. However, in this paper we will not be engaged in resampling, but only sketch out the basic methodology of our approach and its theoretical foundations.
6. In contrast to most theoretical approaches to language, in natural language processing it is far more common to use gradient notions of semantic similarity (cf. the *word space'* model, Schütze 1993; Sahlgren 2006). However, this approach to semantic similarity is mostly applied within a single language, or maximally for the comparison of pairs of languages.
7. "Absolute identity is an abstraction of mathematical thinking. In reality only similarity exists. Identity is strong similarity, is a relative notion. It depends on the sharpness of the senses or of scientific thinking, ultimately on the degree of attentiveness and interest, as to how far, for example, a particular classification is driven." [translation BW]
8. Identity is not the same as indistinguishability. As argued in the main text, being indistinguishable does not necessarily imply identity. In contrast, it is also possible for two meanings to be identical, but still be distinguishable, as is the case of two different forms used for the same referent in reference tracking.
9. Compositional approaches to meaning have the advantage that they can easily distinguish basic from complex meanings. The complexity of a meaning correlates with the length of its paraphrase. However, as pointed out by Dahl (1985: 9), basic meaning need not be viewed intensionally, but can also be viewed extensionally, where "we divide the extension of a term into different regions, one of which we — for whatever reason — look upon as 'basic' or 'primary' with regard to the other" (Dahl 1985: 9). Such an approach is easily compatible with prototype semantics where prototypes are considered to be the best exemplars (see, e.g., Dahl 1985: 4).
10. Cysouw (2010) argues that this underlying map can more generally be conceived as a metric on analytical primitives.
11. Seen as a matter of probability (instead of a universally valid cognitive model), this model of linguistic structure can cope much better with (apparent) exceptions. What traditionally might have been called exceptions are considered here to be simply highly improbable phenomena.
12. This approach to meaning focuses on extension; it is radically usage based. Differences in meaning between lexemes from different languages can only be distinguished to the extent they are represented in the sample in a sufficient number of different contexts. This can lead to unwanted side effects when the sample of contexts is not sufficiently fine grained. For example, it happens to be the case that the contexts for CROSSING in the sample from Mark are all used in situations where crossing takes place by a boat across water. It is thus not possible to distinguish between CROSSING and GO.BY.BOAT in the analysis. So, the selection of the sample of contextually embedded situations determines the resolution of the semantic map.

    However, in this particular case, the semantic similarity between the two meanings might be bigger than commonly believed (CROSSING is path and GO.BY.BOAT is manner in Talmy's 2000 typology). The etymological dictionary of Indo-European verbs (Rix 2001: 472) has the same problem as our semantic map. It reconstructs a root ***per*** 'cross (especially of crossing water)' from which are derived among other things German *fahren* 'go by vehicle', and Greek *poréuomai* 'travel' (see below). It is possible to narrow down the meaning of the verb root in the reconstruction to CROSS and GO.BY.VEHICLE, but it is not possible to reconstruct whether it was a manner or a path verb originally.
13. There are several slightly different kinds of multidimensional scaling. In this paper we use classical multidimensional scaling as implemented by the function `cmdscale` in the software package R (R Development Core Team 2007). See Croft and Poole (2008) for a different kind of MDS applied to typological data.

14. An anonymous reviewer suggests that it would have been preferable to sample Spanish only once. This would certainly be correct if we were trying to approach a stratified sample which is, however, explicitly not the case. Actually the two versions of Spanish are very different, so this bias does much less harm than including several Germanic languages in the sample which behave largely the same. For the effect of sampling in probabilistic semantic maps see Wälchli (2010).
15. We are grateful to Heyka Krause for having made part of the database interpretable for automatic analysis. All data have been coded by BW manually, and he alone is responsible for all errors in the database.
16. Satellites are coded in the database, but our current analysis does not use this information. The current analysis is thus not necessarily ideal, but the simplistic approach chosen here has been the standard approach in motion event research at least since Malblanc (1944) and Tesnière (1959).
17. For an analysis of the multiverb constructions in the current data, see Wälchli (2007).
18. For example, Toaripi, an Eleman language spoken on the coast of Papua New Guinea, has a set of verbs orientating motion events with respect to the beach (Brown 1968): *isai* 'go beachward', *kavai* 'go inland', and *ukavai* 'go toward shore, go inland'. Of these, only *ukavai* is represented in the Toaripi translation of Mark. It occurs only twice, in the passages Mark 1:45 ("Jesus could no more openly enter into the city") and Mark 6:53 ("And when they had passed over").
19. An anonymous reviewer is right in pointing out that narrative 'come' is related to Bühler's (1934: 124) notion of "Deixis am Phantasma", with the crucial difference that it is not always the same deictic words that are used.
20. Mapping their form classes onto the semantic map illustrates that the distribution of the categories is very similar in Mari and Chuvash, but nevertheless they differ in some details. Most importantly, Chuvash has a special 'arrive' verb *šit-* without corresponding verb in Mari.
21. The older Uralicist and Turkologist literature assumes that Chuvash *kaj-* is borrowed from Mari (Räsänen 1920: 244), but the connection of Mari *kajaš* with Finnish *käydä* 'go (roundtrip)' and Livonian *kā'dõ* 'walk' is problematic in that Mari */a/* usually only occurs in loanwords (see, e.g., Rédei 1988: 654).

References

Berkeley, George. 1998 [1710/1734]. *A treatise concerning the principles of human knowledge*. Edited by Jonathan Dancy. Oxford: Oxford University Press.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79: 708–736.

Bowerman, Melissa & Eric Pederson. 1992. Cross-linguistic perspectives on topological spatial relationships. Paper presented at the 91st Annual Meeting of the American Anthropological Association, San Francisco.

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

Brown, Herbert A. 1968. *A dictionary of Toaripi with English-Toaripi index* (Oceania Linguistic Monographs 11). Sydney: University of Sydney.

Bühler, Karl. 1934. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Jena: Fischer.

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.

Chafe, Wallace L. (ed.). 1980. *The pear stories. cognitive, cultural and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Croft, William. 2001. *Radical construction grammar. Syntactic theory in typological perspective*. Oxford: Oxford University Press.
Croft, William. 2003. Typology and universals. 2nd edn. Cambridge: Cambridge University Press.
Croft, William & Keith T. Poole. 2008. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical linguistics* 34(1). 1–37.
Croft, William. 2007. Exemplar Semantics. Draft. http://www.unm.edu/~wcroft/Papers/CSDL8-paper.pdf (accessed 7 January 2011).
Cysouw, Michael. 2007. Building semantic maps: the case of person marking. In Matti Miestamo & Bernhard Wälchli (eds.), *New challenges in typology: broadening the horizons and redefining the foundations* (Mouton de Gruyter Trends in Linguistics Series 189), 225–248. Berlin & New York: Mouton de Gruyter.
Cysouw, Michael. 2010. Semantic maps as metrics on meaning. *Linguistic Discovery* 8(1). 70–95. http://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/2/xmlpage/1/article/346
Cysouw, Michael & Bernhard Wälchli (eds.). 2007. Parallel Texts: Using translational equivalents in linguistic typology. [Special Issue] *Sprachtypologie und Universalienforschung* 60(2).
Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Blackwell.
Dahl, Östen. 2007. From questionnaires to parallel corpora in typology. *Sprachtypologie und Universalienforschung* 60(2). 172–181.
François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change*, 163–215. Amsterdam & Philadelphia: John Benjamins.
Güldemann, Tom. 2008. *Quotative indexes in African languages: A synchronic and diachronic survey* (Empirical approaches to language typology 34). Berlin & New York: Mouton de Gruyter.
Haiman, John. 1985. *Natural syntax*. Cambridge: Cambridge University Press.
Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.
Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language* 2, 211–242. Mahwah, NJ: Lawrence Erlbaum.
Hjelmslev, Louis. 1961 [1943]. *Prolegomena to a theory of language*. Trans. by Francis J. Whitfield. Madison & Milwaukee, WI: University of Wisconsin Press.
de Icaza, Francisco A. 1916. Un falso sistema de investigación literaria — *ir* y *venir*. *Boletín de la Real Academia Española* 3. 75–79.
Kaiser, Georg A. 2002. *Verbstellung und Verbstellungswandel in den romanischen Sprachen.* Tübingen: Niemeyer.
Kaiser, Georg A. 2005. Bibelübersetzungen als Grundlage für empirische Sprachwandeluntersuchungen. In Claus D. Pusch, Johannes Kabatek & Wolfgang Raible (eds.), *Romanische Korpuslinguistik II. Korpora und diachrone Sprachwissenschaft,* 71–83. Tübingen: Narr.
Kaufman, Leonard & Peter J. Rousseeuw. 2005. *Finding groups in data: An introduction to cluster analysis* (Wiley Series in Probability and Statistics). Hoboken, NJ: Wiley.
Kemmer, Suzanne. 1993. *The middle voice*. Amsterdam & Philadelphia: John Benjamins.
Koptjevskaja-Tamm, Maria. 2008. Approaching lexical typology. In Martine Vanhove (ed.), *From polysemy to semantic change*, 3–52. Amsterdam & Philadelphia: John Benjamins.
Krug, Manfred. 2001. Frequency, iconicity, categorization: Evidence from emerging modals. In Joan Bybee & Paul Hopper (eds.), Frequency and the emerging of linguistic structure, 309–335. Amsterdam & Philadelphia: John Benjamins.
Labov, William. 1994. *Principles of linguistic change: Internal factors*. Oxford: Blackwell.
Lehmann, Christian. 2004. Data in linguistics. *The Linguistic Review* 21. 175–210.
Levinson, Stephen C. 2003. *Space in language and cognition: Exploration in cognitive diversity*. Cambridge: Cambridge University Press.

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

Malblanc, Alfred. 1944. *Pour une stylistique comparée du français et de l'allemand: Essai de représentation linguistique comparée*. Paris: Didier.

Malchukov, Andrej L. 2010. Analyzing semantic maps: A multifactorial approach. *Linguistic Discovery* 8(1). 176–198. http://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/2/xmlpage/1/article/350

Malchukov, Andrej, Michael Cysouw & Martin Haspelmath (eds.) (2010). Semantic maps: methods and applications. [Special issue]. *Linguistic Discovery* 8(1).

Marty, Anton. 1908. *Untersuchungen zur Grundlegung der allgemeinen Grammatik und Sprachphilosophie*, erster Band. Halle: Niemeyer.

Mauthner, Fritz. 1982 [1923]. *Beiträge zu einer Kritik der Sprache, erster Band: Zur Sprache und zur Psychologie.* 2nd edn. Frankfurt: Ullstein Materialien.

Myhill, John. 1992. *Typological discourse analysis: Quantitative approaches to the study of linguistic function*. Oxford: Blackwell.

Ogden, Charles K. and Ivor A. Richards. 1966 [1923]. *The meaning of meaning. A study of the influence of language upon thought and of the science of symbolism*, 10th edn. London: Routledge.

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

Ramamurti, Rao Sahib G. V. 1986. *Sora-English Dictionary*. Delhi: Mittal.

Räsänen, Martti. 1920. *Die tschuwassischen Lehnwörter im Tscheremissischen* (Suomalais-ugrilaisen seuran toimituksia 48). Helsinki: Suomalais-ugrilainen seura.

Rédei, Károly. 1988. *Uralisches etymologisches Wörterbuch*, I–III. Wiesbaden: Harrassowitz.

Ricca, Davide. 1993. *I verbi deittici di movimiento in Europa: Una ricerca interlinguistica*. Florence: La Nuova Italia.

Rix, Helmut. 2001. *LIV Lexikon der indogermanischen Verben: Die Wurzeln und ihre Primärstammbildungen.* 2nd edn. Wiesbaden: Reichert.

Sahlgren, Magnus. 2006. *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm: Stockholm University dissertation.

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

Schütze, Hinrich. 1993. Word space. In Stephen J. Hanson, Jack D. Cowan & C. Lee Giles (eds.), *Advances in neural information processing systems 5*, 895–902. San Mateo, CA: Morgan Kaufmann.

Talmy, Leonard. 2000. *Toward a cognitive semantics*, vol. 2: *Typology and process in concept structuring*. Cambridge, MA: MIT Press.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

van der Auwera, Johan. 2008. In defense of classical semantic maps. *Theoretical Linguistics* 34(1). 39–46.

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

Wälchli, Bernhard. 2007. Advantages and disadvantages of using parallel texts in typological investigations. *Sprachtypologie und Universalienforschung*. 60(2). 118–134.

Wälchli, Bernhard. 2010. *Linguistic Discovery* 8(1). 331–371. http://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/2/xmlpage/1/article/356.

Wälchli, Bernhard. 2011. The Circum-Baltic languages. In Bernd Kortmann & Johan van der Auwera (eds.), *The languages and linguistics of Europe: A comprehensive guide* (World of Linguistics 1), 325–340. Berlin & New York: De Gruyter Mouton.