Building semantic maps: the case of person marking

Michael Cysouw

1. Introduction[1]

When comparing languages, a method is needed to deal with variability. Semantic maps are a frequently used method in typology to analyze and display cross-linguistic diversity. In my dissertation on the paradigmatic structure of person marking, I made a semantic map of person marking to summarise some structural aspects of my findings (Cysouw 2001a: 185–187). However, the resulting map was not very satisfying, and I subsequently removed this attempt in the published version of my thesis (Cysouw 2003). In this paper I will revisit the attempt to build a semantic map for person marking, based on the data as discussed in detail in Chapters 3 and 4 from Cysouw (2003). My main conclusion will be that there is no such thing as *the* semantic map for person marking. Instead, various semantic maps are possible, and the traditional kind of semantic map should be seen as just one possible display of cross-linguistic variability. More generally, I will criticize the received view on establishing semantic maps (as summarised in Haspelmath 2003) because frequency of occurrence is ignored in that tradition.

In Section 2, I will first discuss why there is a need for semantic maps and what is the methodological status of this approach in linguistic theory. Following this, in Section 3, I will turn to person marking, which will be the example I will use in this paper to illustrate my arguments. I will first make a semantic map for person marking along the received approach, and then sketch various ways to improve on this by including frequencies of occurrence. Section 4 will outline some possibilities of using similar approaches on different levels of linguistic structure. A summary and an outlook on further possible developments will be given in Section 5.

2. Background and some terminology

The basic impetus for building semantic maps is the variability of linguistic structure among the world’s languages. Elements of a language, be it lexemes, grammatical morphemes, or syntactic constructions, all show language-specific characteristics. It is therefore difficult, if not impossible, to equate two such elements from different languages. For example, the English verb *to fly* is normally translated with the German *fliegen*, but these two lexemes are not identical in all their nuances and idiomatic usages. In the case of such a lexical example, probably nobody would doubt the inherent cross-linguistic variability. However, detailed descriptive work and much cross-linguistic research from the last few decades has shown time and again that the same cross-language incompatibility also exists at all other levels of language structure. This variability poses a problem to large-scale language comparison, because what should be compared with what, when everything is different?

The solution to this problem as used in typological research is to refer to a *tertium comparationis*, normally in the form of a semantically or functionally defined extra-linguistic concept, as the basis of the cross-linguistic comparison. The basic goal of a semantic map is to sketch out the relations between various such *tertia comparationum* as established by the cross-linguistic variability of their structural encoding among the world's languages. As an example, consider the semantic map of indefinite pronouns in Figure 1, as proposed by Haspelmath (1997). In this semantic map, Haspelmath distinguishes nine points of comparison, and the lines connecting these points indicate the relations between them (as established by the cross-linguistic variability). More precisely, when two points *not* connected by a line are both coded by the same pronoun in a particular language, then this semantic map predicts that the points on at least one of the possible connections between these two points will also be coded by that same pronoun.

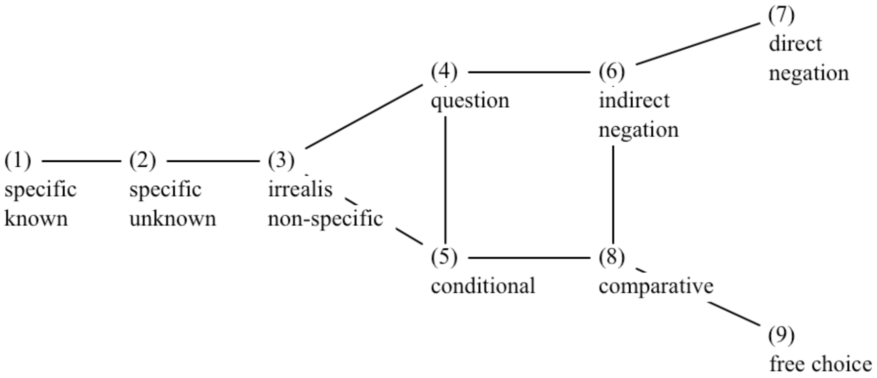

Figure 1. Semantic map of indefinite pronouns (redrawn from Haspelmath 1997)

In his survey of the methodology for building such semantic maps, Haspelmath (2003: 214) uses the term *function* to refer to the points of comparison. I would like to remain somewhat more agnostic about what kind of entities are the basis for a semantic map, and use the term *analytical primitive* instead. An analytical primitive is any concept that is needed for the analysis of a particular set of data. The choice of such primitives is of course informed by the researcher's hypotheses about the structure of human cognition and the social structure of linguistic interaction. However, primitives are primarily meant to be minimal elements of attested linguistic variation. Additionally, I would want to refrain from any claims about them being "universal" primitives of language structure. A particular primitive is used for the analysis of a particular set of data at hand, hence the addition *analytical* primitive (an addition, by the way, that is often dropped for reasons of readability in this paper). It might be the case that such an analytical primitive will withstand refinement by the further development of the field, and thus at one point be considered more than just an analytical tool. However, it is important to keep in mind that we currently only have a rather limited knowledge of the possibilities of human language structure and many of the received opinions about the extent of possible variation will probably have to be revised in the light of yet unknown diversity.

Analytical primitives minimally have to be cross-linguistically interpretable entities. They are the basic elements for cross-linguistic comparison. In contrast, the term *category* will be used here to refer to language-specific elements of linguistic structure. To preclude possible misunderstanding, I will sometimes use the term *language-specific category*, but in general I consider every category to be language specific and simply use the term "category" instead. In the process of typological comparison, the set of analytical primitives is used to describe the meanings or functions of the language-specific categories. For the building of a semantic map, a language specific category is equated to a particular selection of analytical primitives. Often, two different languages apparently have the same category, in the sense that the two categories are described by the same set of analytical primitives. Indeed, on that level of detail, both categories can be considered to be identical, though I expect that by adding more analytical primitives, further differentiation will occur. Cross-linguistic identity is always just a matter of the granularity of analysis.

A semantic map, then, is a structure, based on a set of analytical primitives, that models the variety of categories attested among the world's languages. I would like to stress the usage of the word "model" in this context.

I prefer not to interpret a semantic map as a "theory" of linguistic structure. A semantic map is a model of attested variation, which might, if it turns out to be a good model after many more years of research have passed, be the basis for the formulation of a theory. However, in my opinion we are still far away from any such secure models for them to be called theories. Two subsidiary notions arise when thinking about a semantic map in terms of a model. First, a model depends on the phenomena that one would like to emulate, so there could be different semantic maps for the same set of analytical primitives. A semantic map models attested relations between the various analytical primitives, and thus different results will arise depending on what kind of relations are considered. Second, the accuracy of a model is a factor that can be quantified. In other words, a semantic map should be compared to the data to be modelled, and a measurement should be established how well the model captures the data. In this sense, a semantic map is not just right or wrong anymore, but can be accurate to a certain extent.

Summarising, three things are needed to make a model of the linguistic variation in the form of a semantic map. First, a set of analytical primitives is needed as the basis for cross-linguistic comparison. Second, a set of empirical relations is needed between every pair of primitives. Traditionally, this relation has been either "attested as combined into the meaning of a language-particular category" or "unattested as such". However, I will argue in this paper that such "yes or no" relations might better be replaced by quantitative notions. Finally, equipped with a set of analytical primitives and a set of relations between them, it would be good to have a technique to display any structure in these relations. Note that such a method of display is not necessary to model the variation attested. The analytical primitives and the relations between them already are a model. However, linguists, being just human beings, normally cannot interpret any large table of numbers in a consistent and meaningful way. A good graphical display often tells much more than a thousand numbers. Yet, it is essential to realise that a graphical display is maximally as good as the underlying numbers. In most cases, the graphical display is just a coarse summary of the data, expelling much of the available variation (overgeneralising) or suggesting much more than is actually attested (undergeneralising). Ideally, every graphical display should be accompanied by some measures of accuracy to give an indication of the amount of distortion of the display relative to the data.

In this paper I will discuss various options for the establishment of a semantic map for person marking. Depending on the choice of analytical primitives, on the rationale of establishing relations between the primitives, and on the choice of the graphical display, different maps can be constructed. The choice between the various maps is not one between right and wrong, but one between suitable or unsuitable for a particular goal. Further, it is essential to always explicate what are the underlying assumptions that have been made for a particular semantic map, so that any conclusions drawn from a graphical display are really warranted by the empirical data. Beautiful pictures very easily tell stunning stories to human eyes, but these stories are not necessarily substantiated by the data underlying the graphical display.

3. A semantic map for person marking

3.1. Person marking primitives

Based on a large diversity sample of person paradigms (both in the form of independent pronouns and inflectional person marking), I have argued that at least eight primitives are needed to analyse the world's linguistic diversity of person marking (Cysouw 2003: 72–78).[2] These primitives are summarised in Table 1.[3] The numbers used in the first column of this table are abbreviated *names* for the primitives—they are not a feature-like analysis of their meaning (though the names are intended to have some mnemonic potential). Each person category in a particular language will be analysed as a combination of these primitives. A particular person category mightconsist of just a single primitive, like the English pronoun *I*, which is analysed as primitive "1" only. However, more often than not, a person category will be analysed as a combination of various primitives. For example,

Table 1. Person marking primitives

Primitive	English	Referential meaning
1	*I*	speaker
2	*you*	addressee
3	*he/she/it*	other (i.e. neither speaker nor addressee)
12	*we*	speaker and addressee only ("dual inclusive")
123	*we*	including speaker and addressee ("plural inclusive")
13	*we*	including speaker but excluding addressee ("exclusive")
23	*you*	including addressee but excluding speaker
33	*they*	excluding speaker and addressee

the English pronoun *we* is a combination of the primitives 12, 123, and 13. In the present paper, such combinations of primitives are written using slashes. Thus, the English pronoun *we* is analysed as 12/123/13. All such combinations of primitives as attested in the sample are summarised in Appendix A.

There is by now a long tradition in linguistics to further analyse such person primitives into combinations of person features. Such approaches are inspired by phonological theory, where phonemes are further analysed as bundles of phonological features. There is a wealth of different approaches in phonology to such feature-based analyses, and likewise (often as a direct spin-off) a multitude of them in the realm of person marking (cf. Cysouw 2003: 73, n. 7 and 8 for a quick survey). Most of such feature analyses of person marking are variations on a basic theme using independent features, like [speaker], [addressee], or [plural]. However, any justification for such a feature-based analysis must lie in the observation of morphosyntactic arguments for the presence or absence of each of the features. Most importantly, the set of primitives defined by the presence of a specific combination of features should form a natural class. For example, a feature [speaker] divides the person primitives into two different classes: those containing the speaker {1, 12, 123, 13} vs. those not containing the speaker {2, 3, 23, 33}. A possible argument for such classes could be, for example, the existence of categories 1/12/123/13 and 2/3/23/33, most famously attested in the independent pronouns of Qawesqar, an Alcalufan language from southern Chile (Clairis 1985: 463–464). As these categories are indeed attested as language specific person categories (although not very widespread), there is some evidence for a feature [speaker]. A suitable set of features for the analysis of person marking should be able to model all person categories attested (cf. Appendix A). As far as I can see, none of the feature analyses proposed even comes close to model the wealth of person categories attested. However, a detailed critique of feature-based analyses has to be the subject of another paper. In this paper I will not use such features, but employ the eight analytical primitives as summarised in Table 1 as the basis for the analysis of person marking.

3.2. A traditional semantic map

On the basis of the eight analytical primitives, it is possible to make a semantic map along the lines summarised in Haspelmath (2003). The basis for such a map is a set of person categories which combine more than one of these primitives into their referential meaning. The person categories that were attested in Cysouw (2003) are summarised in Appendix A. I will describe how to make a semantic map on the basis of these person categories.

To make a semantic map, the first categories to look for are person categories formed by combining exactly two primitives. With eight analytical primitives, there are 28 different combinations possible with two of these primitives. Of these theoretical possibilities, fifteen are attested in the sample.[4] These fifteen combinations linking two primitives are minimally needed for a semantic map. These fifteen categories construct a semantic map as shown in Figure 2a.[5] The next step is to control which of the other categories attested are already accounted for by this map. Each category has to be a connected subgraph, meaning that the primitives involved have to be connected by lines. For example, the hypothetical categories 1/2/23/33 and 1/23/33 both are a connected subgraph of Figure 2a, but 1/2/33 is not. Going through the list of categories attested (see Appendix A), five of them turn out not to be accounted for by Figure 2a. These are 2/12/123/13 (5 cases), 2/12/123/23 (4 cases), 12/123/23 (2 cases), 3/12/123/33 (1 case), and 12/123/33 (1 case).[6] Some lines have to be added to account for these attested categories. There are various equivalent possibilities to add connections to account for these person categories. For example, the category 12/123/23 can be accounted for by either adding the connection 12–23 or 123–23. In this situation, Haspelmath (2003) does not present a principled way to decide between these alternatives.[7] The intuition of the researcher evaluating the resulting graph has to decide. For example, I might propose to add the connections 2–12, 123–23, and 123–33 as shown in Figure 2b, based on reasons of visual symmetry.[8] This is a semantic map for person marking that fits the data in Cysouw (2003).

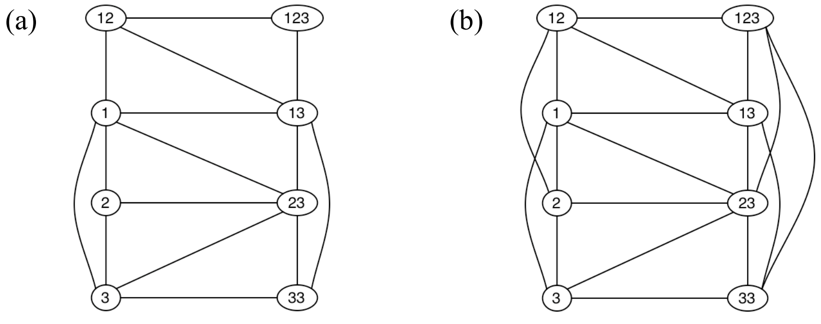

Figure 2. Semantic maps of person marking

3.3. Including frequencies

There are two problems with such semantic maps. First, the boundary between attested and unattested is given very high prominence, or, put more general, the differences in frequency of attestation are ignored. Second, there are many connected subgraphs that are predicted by a semantic map, but that are not attested in the data. In most slightly more complex maps, the ratio of predicted to attested categories quickly becomes much too large to count as a good model.

Starting with the first problem, the semantic map of person marking shown in Figure 2b indeed accounts for all categories attested. However, it is questionable whether the missing lines in the map are really unattested in human languages, or only accidentally not found in the languages examined in Cysouw (2003). Specifically searching for such cases not yet accounted for, it did not take much time to come up with various "counter" examples. For example, in Estonian, verbs in the past have the same form for the second singular and the third plural, so this is a 2/33 category not yet accounted for by the semantic map (Erelt, Erelt, and Ross 2000: 226). In Daga, the past forms of the class A verbs do not distinguish between the first singular and the third plural (Murane 1974: 52–54), so this is a 1/33 category also not accounted for by the semantic map. Finally, in the Diola-Fogny "short version" of the person prefixes, there is no distinction between the third singular and the exclusive (Sapir 1965: 90). This is a 3/13 category also not accounted for by the semantic map. Even after these links are added, there are still a few combinations not attested. However, I do not see any principled reason why these combinations should be absent from the world's linguistic diversity.[9]

The central point is that there does not appear to be a crucial difference between categories that are unattested, and categories that are only attested in very few languages. The difference between these two situations most likely reflects incidental effects of the language investigated, and not any preference of human language structure. In a different sample, it is very probable that other rare categories might be found. Still, this rather superficial difference between attested and unattested is crucial for the establishment of semantic maps in the tradition as summarised by Haspelmath (2003). In contrast, the fact that some person categories are extremely widespread, while others are exceedingly rare is not of importance for the building of a semantic map. Each exemplar, however common or rare it might be, is deemed equally important. Yet, the distinction between common and rare types is in most cases extremely robust, and to a large extent independent of the details of the sample used. Common types will normally be found widespread in whatever sample is used. This difference between common and rare categories seems to be a much more important fact to be modelled than the difference between rare and non-existing categories.

A straightforward solution for this problem is to draw lines in a semantic map with a thickness proportional to its frequency of occurrence.[10] In Appendix B, the frequencies for every pair of primitives are given. In most cases, such a frequency is the sum of occurrences of more than one category. For example, the primitives 1 and 2 have eight co-occurrences. This number is the sum of the frequencies of the four categories that include both the primitives 1 and 2, namely the categories 1/2 (3 cases), 1/2/3 (3 cases), 1/2/12/123/13/23 (1 case), and 1/2/12/123/13/23/33 (1 case). A semantic map using the frequencies from Appendix B to determine the thickness of the lines is shown in Figure 3. This picture gives an informative view on the relative importance of the possible connections between the primitives. However, already with the eight primitives in this example, the semantic map becomes rather messy. When more primitives are added, a display will only become less appealing and more difficult to interpret.

The second problem with the traditional semantic map is that there are many categories that are predicted by this model, but not attested in the current data. For example, the semantic map in Figure 2b predicts the existence of a category linking the primitives 1/2/3/12, though this category is currently unattested. This is not necessarily a bad thing, as a good model always predicts a few things not yet encountered. Such predictions can guide future research. However, the number of predictions should not be exceedingly large in relation to the explained data. As models are in general only approximations of reality, they will always show some excess (catego

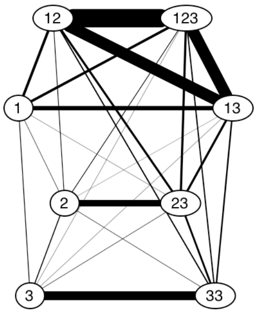

Figure 3. Semantic map of person marking informed by frequency of attestation

ries predicted that are unattested) and some deficiencies (categories attested, yet not covered by the model). I will here use the term “coverage” for the number of attested categories captured by the model devided by all attested categories, i.e. a coverage of 100 % means that all categories attested are included in the predictions of the model. Additionally, I will use the term “accuracy” to refer to the fraction of attested categories among all categories that are predicted by the model, i.e. a 100 % accurate model only predicts those categories that are attested.

In building semantic maps one will have to find a balance between coverage and accuracy. The tradition of building semantic maps in linguistic typology prefers high coverage to high accuracy. For example, Haspelmath’s (1997) semantic map for indefinite pronouns (cf. Figure 1) has a coverage of 100 %, but it actually predicts the existence of 105 different categories of which only 39 are attested, i.e. an accuracy of only 35 % (cf. Cysouw 2001b). Looking at it this way, Haspelmath’s model is rather inappropriate. To investigate the balance between coverage and accuracy, one should ideally investigate a large number of semantic maps, and look for cases that strike a good balance between the two.[11]

For the current case study of person marking, I investigated a few semantic maps that maximise coverage relative to the number of lines (see Appendix C). The balance between coverage and accuracy for these semantic maps is shown in Figure 4. For coverage, I counted the number of categories accounted for by the semantic map, and divided this through the total number of categories attested in the sample. For these calculations, I counted category *tokens*, i.e. I took the frequency of occurrence of each category into account. For accuracy, I counted the number of categories accounted for by the model, and divided this through the number of categories predicted by the model. Here I counted category *types*, i.e. I did not take the frequency of occurrence into account. The reason is that I do not have any ground for assessing the frequency of unattested types. The points in Figure 4 represent different semantic maps, with an increasing number of lines going from left to right.[12] The first few lines that are added raise the coverage without leading to any inaccuracy. However, starting with the fifth line, some categories are predicted that are not attested. Subsequent lines that are added still improve the coverage, but the accuracy rapidly declines. Two different maps in this range are presented in Figure 5. In Figure 5a the map with five lines is shown with a reasonable high coverage (93.1 %) and only very few unattested predictions (accuracy of 93.8 %. There is actually only one unattested prediction, viz. 1/123/13). This can be interpreted as a fine, though slightly conservative model. In Figure 5b, a map is shown in with a very high coverage (over 99 %), but this map predicts 120 categories of which only 34 are attested (accuracy of 28.3 %).[13] This can be seen as a rather courageous model. A good model will be somewhere in the range between these two extremes.

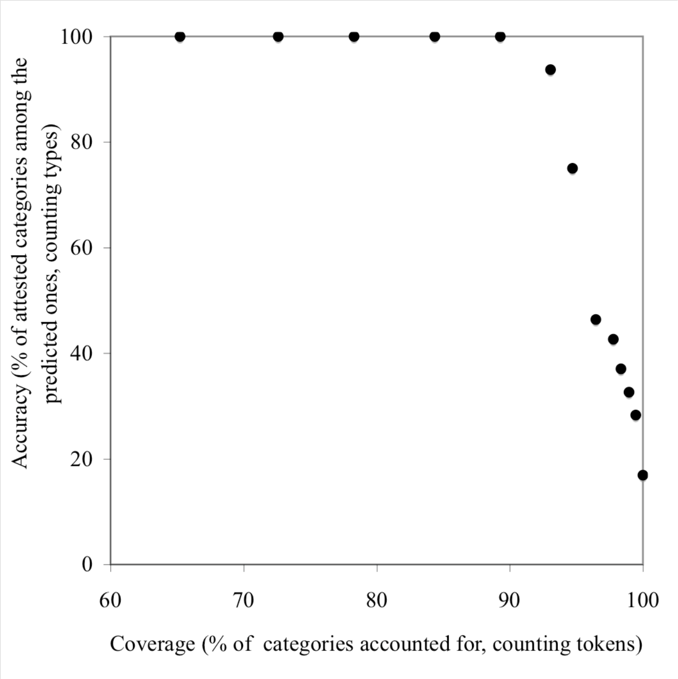

Figure 4. Searching for an optimum between coverage and accuracy

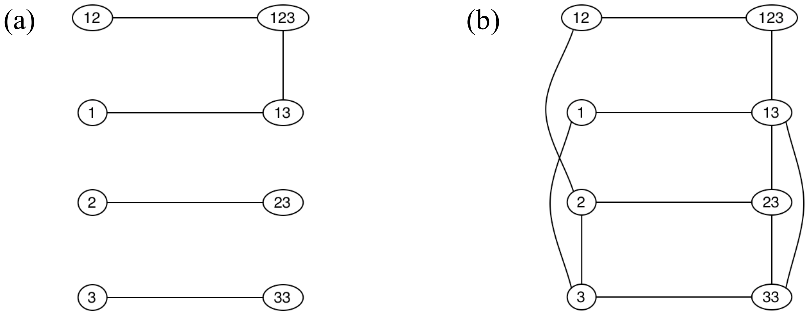

Figure 5. Semantic maps with a different balance between coverage and accuracy

3.4. Multidimensional scaling

A different approach to including the frequencies of attestation is to interpret the frequencies of co-occurrence (as summarised in Appendix B) as a measure of similarity. The higher the number of co-occurrences of two primitives, the more similar are these two primitives. The internal structure of such similarity-matrices can be displayed using multidimensional scaling (MDS). Glossing over the mathematical details, the idea behind an MDS is that two objects that are similar are placed close to each other, and objects that are less similar are placed further away from each other. Intuitively, an MDS display, like the one shown in Figure 6 for the eight person primitives, can be interpreted like a Euclidean space, in which the distances between the primitives are indicative of their difference. The dimensions of an MDS display are nameless, and only indicate some general mathematical notion of similarity.

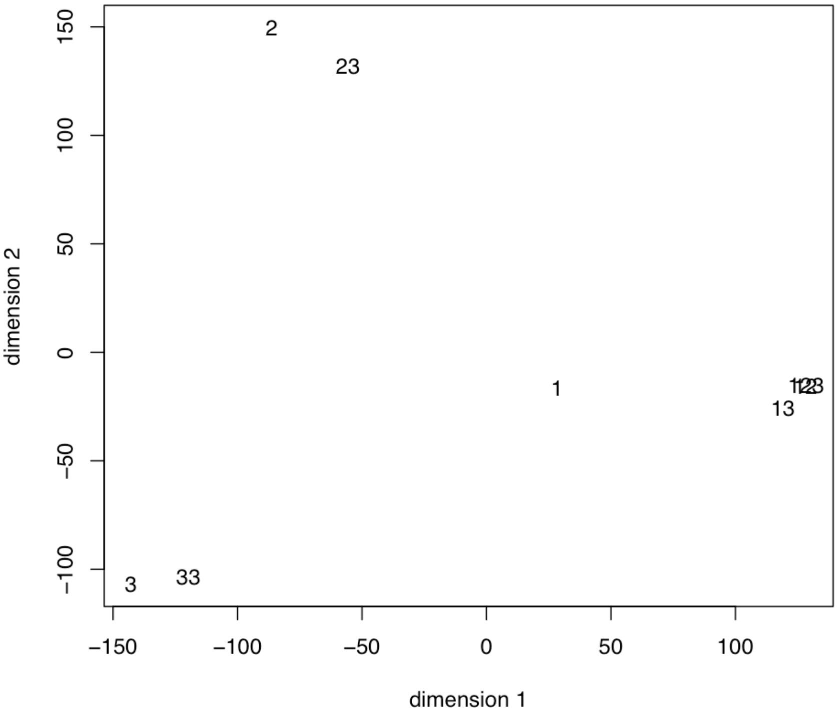

Figure 6. Multidimensional scaling of the person marking primitives

Although a display as shown in Figure 6 is extremely helpful when trying to make sense of a messy set of data, it is important to realise that an MDS represents a strong reduction of the available information. I would therefore suggest not to consider an MDS to be an improvement over the traditional semantic map (as proposed by Croft and Poole 2004). Without diving into the mathematical details of an MDS, let me explain this problem. When there are only three objects, with distances measured for every pair, then it is always possible to place these three objects on a two-dimensional plane in such a way that the pairwise distance between the points on the plane is exactly proportional to the measured distance (viz. three distances determine a triangle). However, with four objects this is not always possible. In most cases, a third dimension is needed to display the distances to the fourth object exactly. In principle, an extra dimension is needed for every object to be added. So for eight objects, as in the current case of person marking, seven dimensions would be needed. Now, the mathematical procedure called MDS tries to show as much as possible of the actual distances in only two dimensions. To do so, some of the distances have to be changed a bit. The MDS searches for the minimal amount of changes needed to display all objects in two dimensions. In a sense, this is comparable to removing some of the thinner lines from Figure 3. The selection of only two dimensions for an MDS is necessarily a reduction of the actual variation, focussing on the oppositions with the highest frequency. Of course, it would be possible to display a three-dimensional MDS, but that would still omit the dimensions four to seven. A display like Figure 6 is very useful to get some insight into the major dimensions of variation in a particular dataset, but it does not suffice as a model for the cross-linguistic variation, because there is an arbitrary cut-off point of data-reduction as determined by the dimensionality of display.[14]

4. Up one level: Mapping categories

Whatever semantic map will be considered most suitable for the problem at hand, it remains to be remembered that a semantic map (only) serves one particular function, namely to model categories on the basis of a set of analytical primitives. The question how these different categories relate to each other is not answered by establishing a semantic map. Investigating the relation between the categories themselves is in principle a straightforward extension of the methods discussed previously. It is possible to consider the categories themselves as a kind of analytical primitives, and make a semantic map linking these categories to each other. Such an analysis can be considered a "second-order" semantic map. However, I would like to restrict the name semantic map for an analysis based on a set of (semantically based) analytical primitives. The higher order maps to be discussed in this section are analyses that take the categories themselves as a basis. Therefore, I would propose the name CATEGORY MAP for such a display, showing the interrelation between language specific categories.

When making a category map, immediately the problem arises that there normally are very many different categories. In the present case of person marking, there are only eight analytical primitives, but 43 different categories. Searching for an optimal graph-structure already is problematic with eight primitives, the more so with 43, considering that the number of possible graphs rises exponentially with the number of primitives considered (see Footnote 11). For that reason, I will use multidimensional scaling here to investigate the relation between person categories. Remembering the provisos made in the previous section, the MDS displays to be discussed shortly are only attempts to find some structure in a large set of messy data. They are not intended to be models of cross-linguistic variation (let alone a theory).

Trying to relate the different language specific categories to each other, the question arises what kind of relation to consider. More practically, the question is how to measure the similarity between two different language-specific categories. A straightforward approach is to take the relative overlap of primitives. For example, the categories 2/23 and 12/123/13/23 share one primitive (viz. 23) out of a maximally possible overlap of two primitives (viz. 2 and 23), giving a relative overlap of 0.5. This relative overlap can be computed for every pair of primitives. The MDS based on these distances is shown in Figure 7. The first dimension (depicted horizontally) distinguishes the categories including all the reference of English *we* (12/123/13) to the right from all categories not involving any kind of 'we' to the left, with all categories marking some kind of inclusive/exclusive distinction in the middle. The second dimension (depicted vertically) is somewhat more difficult to characterise, though is appears to separate categories including primitive 1 ('speaker') to the top from categories including primitive 33 ('third person plural') to the bottom, with everything else in between.

A more interesting measure of category similarity is based on the co-occurrence within paradigms. Person markers typically form a paradigm of

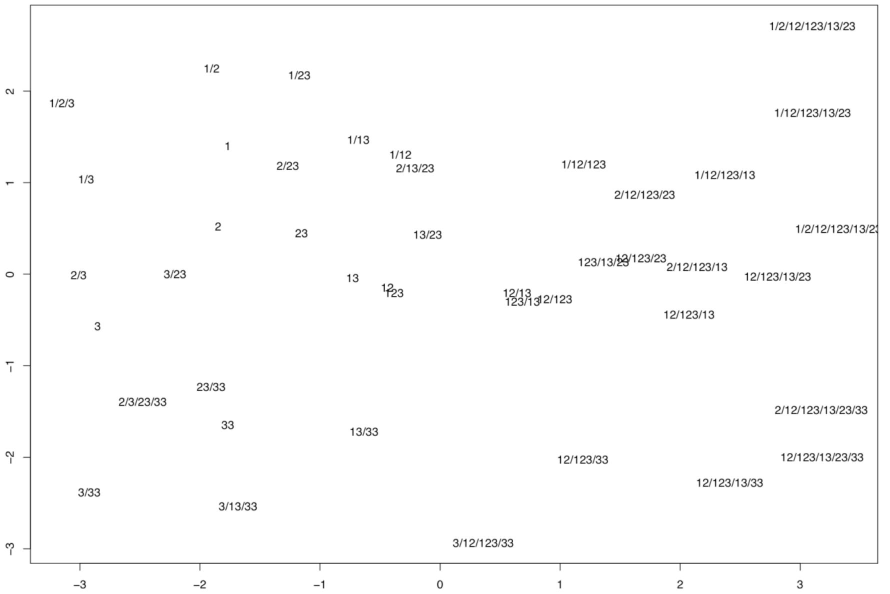

Figure 7. MDS of person categories. The underlying similarity between the categories is established by the relative number of shared primitives

syntagmatically related forms. Now, a straightforward measure of similarity between two categories is the number of paradigms that contain both categories. The idea underlying this measure is that if two categories regularly co-occur in a paradigm, this might indicate some connection between these categories. For example, in the present collection of paradigms, there are 72 paradigms in which both the categories 3/33 (i.e. number neutralization in the third person) and 2/23 (i.e. number neutralization in the second person) occur. The fact that this combination occurs so often seems to indicate that these two categories are linked (cf. there is might be a tendency to have number neutralization throughout various persons).

The frequencies of paradigmatic co-occurrence for every pair of categories is computed, and the MDS based on these frequencies is shown in Figure 8. This display shows some aspects of the paradigmatic structure of person marking (cf. the title of Cysouw 2003). The first dimension (depicted horizontally) distinguishes the categories consisting of only one of the eight analytical primitives to the left from the categories showing number indifferentiation to the right. The second dimension (depicted vertically) makes an interesting separation between the inclusive at the bottom

(12/123), then a group of categories involving mixes of inclusive or exclusive with other primitives in the middle, and a rather mixed group at the top. This mixed group contains both categories including all reference of English *we* (12/123/13) and categories mixing up person categories that are not including any 'we' (e.g. 1/3, 1/23, or 2/3/23/33). That these two kinds of categories belong together is also observed in Cysouw (2003)—though without such graphical display—under the name of *pure person* marking (Cysouw 2003: 163). Paradigms that "purely" mark person have at least some inclusive/exclusive distinction. Only when this distinction is not made in a person paradigm, then other person confusions might happen in a paradigm. This group of "non-pure" person markers is the one found to the top of the MDS in Figure 8. It is highly stimulating to find this inductively established regularity also in the mathematically established category map as shown in Figure 8.

These two category maps are only examples to illustrate the principle of extending the idea of semantic maps to higher structural levels. They are proposed as showing some new possibilities of mapping relations between

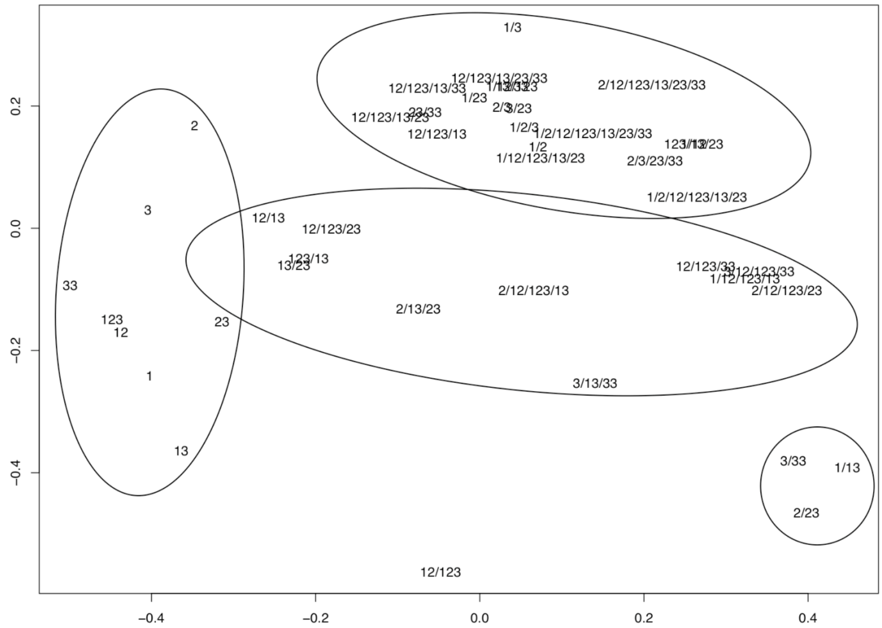

Figure 8. MDS of person categories. The similarity between the categories is established by counting the number of paradigmatic co-occurrences

linguistic structures, indicating directions for further research. Further possibilities can be found in the establishment of similarities and in the elements of comparison. First, there will surely be many more possibilities to establish similarities between categories (I have discussed only two possibilities here), often depending on the kind of categories investigated. Second, there are also other kinds of higher-level structural comparisons possible. For example, it seems a worthwhile project to establish a measure of similarity for the complete structure of paradigms, and establish a "paradigm map" showing the relation between paradigms (cf. Cysouw 2003: 245–294 for a non-mathematical attempt at establishing a paradigm map).

5. Conclusions

In this article, I have sketched a few possible directions that the analysis of semantic maps can take once larger amounts of data become available. The methods to investigate large sets of typological data are still in their infancy, and these proposals are meant to be a step towards slightly more sophisticated typological modelling of large datasets.

The main problem with the received approach to construct semantic maps is that it will probably not work when larger sets of data are considered. Using the approach to building semantic maps as succinctly summarised by Haspelmath (2003), it is very well possible that the resulting maps do not stand the test of additional data. The reason is that every new possibility attested has to be accounted for. Consequently, many rare phenomena, which will surely be found once more data is considered, will force correction of the semantic map. Because traditional semantic maps focus on *possible* human language structure, I would even predict that in the end most (if not all) of the possible lines linking two primitives will be needed to describe the wide variety of structural possibilities that the human language capacity can deal with. Much (if not everything) is possible in human language structure, though not everything is equally probable. It might be wise instead to focus more on modelling *probable* human language structure by taking into account the frequencies in which particular combinations are attested.

Interpreting a semantic map as a model for linguistic variation alleviates some of the pressure often put on a semantic map. The map does not have to be perfect. It should have a good coverage, but also a high accuracy. The received approach to semantic maps favours coverage above accuracy, because of the focus on possible language structure. When changing the view to model probable language structure, the option arises to search for a balance between coverage and accuracy. Both measures should be high, though neither has to be perfect.

Finally, it is important to realise that a model only helps us understand the phenomena under consideration. A single model will never be able to capture all observations within a particular domain. Specifically for the domain of person marking, I have discussed how to investigate the cross-linguistic diversity of categories by making a semantic map of the eight person primitives (Section 3). Further, I have given some first hints as to the investigation of paradigm structure taking the 43 person categories as the primitives of a category map (Section 4). However, there are more levels of analysis possible. For example, it is possible to analyse the primitives by looking at even smaller elements like features (cf. the discussion in Section 2), and it is possible to investigate the relation between the structure of whole paradigms as attested among the world's languages. Combining all these possibilities, there are at least four different levels of analysis in which methods like the ones discussed in this paper can be of service: features to analyse primitives; primitives to analyse categories; categories to analyse paradigms; and paradigms to analyse languages. Taking any one level as the starting point, it is possible to model the variation at the next level. However, I think it is a mistake to bypass levels and, for example, try to parameterize the variation among whole languages on the basis of a set of features. One should be taking one step at a time. Modelling is a modest business.

Appendices

Appendix A: Person categories

With eight analytical primitives, there are theoretically 2^8-1-1= 254 different person categories possible (minus one for taking none of the primitives, and minus one for taking all primitives; these combinations do not make sense as person categories). In total, there are 43 person categories attested. First, all eight primitives are attested individually as person categories. Further, there are in total 35 different combinations of the basic eight person primitives attested in the 325 person paradigms (from Cysouw 2003: Ch. 3 and 4). Only these 35 combinations of primitives are shown in this appendix, ordered by frequency. The most frequent combinations are easily interpretable referentially.

Categories	Approximate meaning	Freq.	Categories	Freq.
3/33	‘third’	125	123/13	3
12/123/13	‘first plural’	100	1/2	3
12/123	‘inclusive’	97	1/2/3	3
2/23	‘second’	84	12/13	2
1/12/123/13	‘first’	35	13/23	2
1/13	‘exclusive’	29	3/23	2
12/123/13/23	‘non-third plural’	18	12/123/23	2
23/33	‘non-first plural’	17	1/12/123/13/23	2
12/123/13/33	‘non-second plural’	11	123/13/23	1
1/3	‘non-second singular’	10	13/33	1
2/3	‘non-first singular’	7	1/12	1
2/3/23/33	‘non first’	6	1/23	1
3/13/33		5	12/123/33	1
2/12/123/13		5	1/12/123	1
12/123/13/23/33		5	3/12/123/33	1
2/13/23		4	1/2/12/123/13/23	1
2/12/123/23		4	2/12/123/13/23/33	1
			1/2/12/123/13/23/33	1

Appendix B. Frequencies of pairwise co-occurrence of person primitives

For every pair of primitives, the total number of categories was counted in which both primitives occurred.

	2	3	12	123	13	23	33
1	8	13	41	40	68	5	1
2		16	12	12	4	101	8
3			1	1	5	8	137
12				286	181	34	20
123					184	35	20
13						35	24
23							30

Appendix C. Evaluating different semantic maps as to coverage vs. accuracy

The first two columns of this appendix describe a selection of semantic maps for person marking, adding lines subsequently. The third and fourth column described the coverage of these semantic maps, counting tokens. In total, there are 591 combinations of person primitives attested (i.e. the sum of the frequencies as listed in Appendix A) Further, there are 1109 occurrences of categories that only consist of one single primitive. In total, there are thus 1700 person categories in the present sample. The coverage is the number of categories accounted for divided by 1700. The last three columns describe the accuracy of the semantic maps. Out of the 254 possible categories only 43 are attested. The accuracy for the complete graph is thus 43/254 = 16.9 %. Likewise, the accuracy is established for all other semantic maps considered.

No. of lines	Line added	Categories accounted for (tokens)	Coverage (%)	Categories accounted for (types)	Categories predicted (types)	Accuracy (%)
0		1109	65.2	8	8	100.0
1	3—33	1234	72.6	9	9	100.0
2	12—123	1331	78.3	10	10	100.0
3	13—123	1434	84.4	12	12	100.0
4	2—23	1518	89.3	13	13	100.0
5	1—13	1582	93.1	15	16	93.8
6	13—23	1610	94.7	21	28	75.0
7	23—33	1640	96.5	26	56	46.4
8	13—33	1662	97.8	29	68	42.6
9	1—3	1672	98.4	30	81	37.0
10	2—3	1682	98.9	32	98	32.7
11	2—12	1691	99.5	34	120	28.3
…	...	…	…	…	…	…
28	All lines	1700	100.0	43	254	16.9

Notes

1. I thank (in alphabetical order) Balthasar Bickel, Martin Haspelmath, Elena Maslova, Matti Miestamo, Nicoletta Puddu and Bernhard Wälchli for their helpful comments on earlier versions of this paper.
2. Two further primitives of person marking that might be considered are "choric we" (i.e. a group of only speakers, speaking in unisono) and "present audience" (i.e. a group of addressees only, being addressed together). Although these are conceptually sensible primitives, there is currently no evidence known among the world's languages that such primitives are needed to analyse the person markers in the world's linguistic diversity (cf. Cysouw 2003: 72–78; Simon 2005).
3. These primitives include both singular and plural notions. One of the central claims of Cysouw (2003) is that the *nominal* notion "plural" has no place in the analysis of person marking (cf. Daniel 2005, who finds only 7% of the world's independent pronouns to be composed of a person stem with a *nominal* plural affix). In contrast, categories like dual, trial, or paucal are considered to be number categories in the realm of person marking. For reasons of space, these number categories are disregarded in this article.
4. The variability attested for person marking might seem large, and maybe other domains of linguistic structure are more constrained. However, there are not many domains of linguistic structure for which currently the possibility exists

to perform an investigation with such a strong emphasis on diversity as I have been able to do for person marking. I have included every "odd" structure that I could find, thereby increasing the diversity to match the variability that would otherwise only be found in a much larger sample (impressionistically somewhere in between 1,000 and 2,000 languages). I expect that such extremely large samples will also lead to large variability in other domains of linguistic structure.

5. The layout of the primitives in this figure is inspired by the minimal/augmented person paradigm in which the dual inclusive (12) is aligned with the singular categories (cf. Cysouw 2003: 85–90). However, this aspect of the depiction is relatively unimportant, as for a semantic map only the graph structure of the connections counts.
6. None of these categories is particularly common among the world's languages. Also note that, except for the first, these categories are combinations of the inclusive (12/123) with either the second person plural (23) or third person plural (33). Even in larger samples of the world's linguistic diversity such categories are only rarely found. However, in an in-depth investigation of these quirks, Cysouw (2005) confirmed that their existence is a robust phenomenon.
7. Already in this rather simple case there is no principled way to decide between alternative connections in the semantic map, though there are many possible approaches to make this choice more constrained. Such procedures will become even more important in datasets that do not have as many pairwise connections between the primitives as are found in the present sample. A relatively straightforward approach would be to prefer connections that are needed for larger sets of cases in the data.
8. The argument of visual symmetry is of course completely dependent on the layout of the primitives. A different layout will invoke different pleasing visual effects. This argument should thus be taken only as exemplary for the many ad hoc arguments that might lead a researcher to propose a particular semantic map.
9. There is one consideration that I have not included here, and that is the argument of "incidental" categories. There is a recurrent argument to be found in the linguistic literature that some categories are to be considered incidents of history, and that such incidents do not have to be explained by a theory of linguistic structure. A commonly cited example of such an incident is a phonological merger leading to the synonymy of two erstwhile different categories. I consider it bad practice to disregard such cases in the collection of data, as the argument of incidentality might often be used to explain away examples that do not fit the theory to be proposed. Even when a merger can explicitly be shown to have taken place, this still raises the question why the resulting syncretism is not immediately disambiguated. Concerning linguistic structure, the question is not what the origin of a structure is, but how likely it is that the re-

sulting situation will occur. If a particular merger is indeed incidental, then there should be a low chance of occurrence of the resulting structure in a typological sample. Taking frequencies into account will thus implicitly delimit the influence of real incidental structures—without there being any need for a clear-cut decision on what should count as incidental, and what not.

10. It is a big problem exactly which frequencies should be deemed relevant to make a semantic map. As any set of empirically collected frequencies depend on the sample of languages chosen, so the whole issue of sampling in typology comes up at this question. This discussion will not be taken up here. The sample used in the present example of person marking can best be called a convenience-diversity sample, and might thus not be really applicable to an investigation of universal patterns in human language. However, the sample is still suitable to exemplify the possible usage of frequencies, whatever their meaning might be.
11. Ideally, all possible semantic maps should be investigated. However, this number will quickly get very high. For example, in the current case of person marking with eight primitives there are 28 possible lines. Each of these lines can be present or not in a semantic map, giving a total of $2^{28} \sim 2.7 \times 10^{8}$ possible semantic maps. Dealing with such large search spaces is difficult, though in computer science there are various kinds of sophisticated search algorithms available that could be used to approximate optimal solutions.
12. Only shown here are the addition of the first eleven lines. So, there are thirteen points in the figure: one for the model without any lines at all, one for the model with all 28 lines, and eleven points for the models in between, corresponding to the subsequential addition of one line. No models with 12 to 27 lines are shown.
13. Note that the map as shown in Figure 5b could easily be given a visually much more pleasing layout without crossing lines. However, I have not changed the layout for reasons of comparability with the other maps discussed in this article.
14. The measure for the amount of data-reduction of an MDS is called the "stress" (normally given in percentages). The lower the stress, the better the display reflects the underlying distances. Figure 6 has a stress of 12 %, which is actually not bad at all. For three dimensions, the stress even falls below 1 %, so that is actually a rather good model for the case of person marking.

References

Clairis, Christos
1985 *El Qawesqar: Linguïstica Fueguina, Teoria y Descriptión.* Valdivia: Universidad Austral de Chile.

Croft, William, and Keith T. Poole

2004 Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. Unpublished manuscript. Available online at <http://www.unm.edu/~wcroft/WACpubs.html>.

Cysouw, Michael

2001a The paradigmatic structure of person marking. Ph.D. dissertation, Radboud University Nijmegen.

2001b Review of *Indefinite Pronouns* (Haspelmath 1997). *Journal of Linguistics* 37/3: 99–114

2003 *The Paradigmatic Structure of Person Marking.* (Oxford Studies in Typology and Linguistic Theory). Oxford: Oxford University Press.

2005 Syncretisms involving clusivity. In *Clusivity: Typology and Case Studies of the Inclusive-Exclusive Distinction*, Elena Filimonova (ed.), 73–111. Amsterdam: Benjamins.

Daniel, Michael

2005 Plurality in independent personal pronouns. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 146–149. Oxford: Oxford University Press.

Erelt, Mati, Tiiu Erelt, and Kristiina Ross

2000 *Eesti Keele Käsiraamat.* Tallin: Eesti Keele Sihtasutus. Available online at <http://www.eki.ee/books/ekkr/>.

Haspelmath, Martin

1997 *Indefinite Pronouns.* (Oxford Studies in Typology and Linguistic Theory). Oxford: Clarendon.

2003 The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, Michael Tomasello (ed.), 211–42. Mahwah, NJ: Erlbaum.

Murane, Elizabeth

1974 *Daga Grammar: from morpheme to discourse.* (Summer Institute of Linguistics Publications in Linguistics 43). Oklahoma: Summer Institute of Linguistics.

Sapir, J. David

1965 *A Grammar of Diola Fogny: A language spoken in the Bass-Casamance region of Senegal.* (West African Language Monograph Series 3). Cambridge: Cambridge University Press.

Simon, Horst

2005 Only you? Philological investigations into the alleged inclusive-exclusive distinction in the second-person plural. In *Clusivity: Typology and Case Studies of the Inclusive-Exclusive Distinction*, Elena Filimonova (ed.) 113–50. Amsterdam: Benjamins.