Inducing Semantic Roles

Michael Cysouw
Research Unit ‘Quantitative Language Comparison’
Ludwig Maximilians Universität München
cysouw@lmu.de

Abstract

Instead of defining semantic roles on the basis of the interpretation of lexical predicates, I will show that it is possible to induce semantic roles from the usage of case-like markers across a wide variety of languages. The assumptions behind this proposal are, first, that semantic roles are strongly contextually determined and, second, that similarity in coding of contextual roles across many different languages shows which contexts evoke the same (or better: very similar) semantic roles. This approach to the investigation of semantic roles will be exemplified by an investigation of case-like marking in a parallel text across a sample of fifteen languages. On this basis, a semantic map of contextual roles can be established, and it will be shown that higher level abstractions, like semantic roles or even macro-roles, can be statistically derived from this diversity of marking across many languages. Further, a typology of alignment systems can be derived statistically.

1. Introduction

The notion of SEMANTIC ROLES or THEMATIC RELATIONS (two terms which I will treat synonymously here for ease of discussion) have a long tradition in linguistic analysis (cf. Blake 1930; Fillmore 1968 for some early discussion). A semantic role can be considered an intermediate level of abstraction in between highly abstract proto-roles like agent or undergoer (Dowty 1991) and concrete verb-specific semantic roles, like the giver, knower or walker. VanValin (2004: 64) visually displays these levels of abstraction as a hierarchical clustering, in which lexically specified semantic roles cluster into a smaller set of thematic relations, which in turn cluster into a few macroroles.

In this paper I will propose to add an even more concrete level below this hierarchy of roles, namely CONTEXT-SPECIFIC ROLES. The basic idea behind these roles is that even low-level roles like the giver or knower are abstractions over the actual occurrences of giving and knowing in context. The ultimate basic entity is a specific person in a concrete context in which giving is taking place, the details of which are of course different in each concrete context. The verb-specific role of ‘the giver’ is a cluster of all many such different (though mostly highly similar) concrete contextual roles.

Further, I will argue that it is possible to induce higher level of role-abstractions (alike to semantic roles or proto-roles) from the diversity of overt marking across a wide variety of languages. Basically, the contextual distribution of case-like markers across a wide variety of languages allows for the specification of a metric on the contextually-specified roles. This metric can be interpreted as a semantic map of contextual roles (Cysouw 2010). By using various kinds of statistical clustering, higher level roles can be induced from this underlying semantic map.

In this paper, I will first summarize some of the underlying assumptions on which this kind of research is based (Section 2). I will then describe the data that has been used for the analysis (Section 3), followed by an analysis of the contextual roles in this data (Section 4). Finally, I will discuss the analysis of the alignment patterns of the languages investigated, arguing that it is also possible to statistically derive a language typology from the same data (Section 5).

2. Using cross-linguistic variation to approach semantics

The theoretical assumptions on which the research in this paper is based, EXEMPLAR SEMANTICS and the ISOMORPHISM HYPOTHESIS, are described in more detail in Wälchli & Cysouw (2011) and will only be summarized here. First, the isomorphism hypothesis claims that 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. Individual languages will of course dramatically diverge from this general pattern in their coding of specific meanings (i.e. highly similar meanings might be formally distinguished in a specific language, while highly divergent meanings might be coded identically). However, by averaging over the structures of many languages, these idiosyncratic patterns will vanish among the cross-linguistically recurrent patterns. The isomorphism hypothesis thus implies that cross-linguistically recurrent formal similarities will be indicative of the meanings expressed. In this interpretation, cross-linguistic variation of formal encoding provides evidence for semantic similarity of the encoded events.

Second, EXEMPLAR SEMANTICS is a cover term for all approaches to semantics in which exemplar meaning is considered more fundamental than the meaning of abstract concepts. The assumption is that individual utterances have a very concrete meaning, strongly depending on the context in which they are uttered. The ‘overall’ meaning of any linguistic formative (be it a lexeme, morpheme, or construction) is only an coarse summary of the individual, and highly specific, meaning each individual occurrence of the formative has in each specific context of utterance.

Individual expression as they occur in their context of utterance are thus considered to be the ultimate exemplars. The context of an expression can be defined in general as the spatio-temporal surrounding of an individual expression. This notion of context is deliberately left rather vague here because its precise delimitation depends on the practical implementation in a specific empirical study. The spatio-temporal surrounding of an expression can be defined as the sentence in which the expression occurs, or as the complete text around the expression, or it can even include the socio-cultural setting in which the expression is uttered.

Translated to the concrete case of semantic roles, the assumptions behind the current investigation are the following. First, this study is exemplar-based in that semantic roles are considered to be strongly contextually determined. To a large extent, it is the lexical predicate that determines the roles, but other contextual factors will further specify the precise role a participant takes in any situation. In effect, each participant in context is assumed to be a different *contextual* role. Second, isomorphism is assumed to be the empirical basis of this investigation. The coding of contextual roles across many different languages shows which contexts evoke the same (or better: very similar) roles. Basically, given two participants in different contexts, the more often these two participants are marked identically in language after language, the more similar the contextual roles will be. This similarity can be used to induce higher level abstraction, like semantic roles or proto-roles.

3. The data: case-like marking in parallel texts

The approach to the investigation of semantic roles as described in the previous section will be exemplified by a study of overtly marked nouns in a parallel text across many languages. Strictly for reasons of convenience I will use religious brochures from watchtower.org, translation of which are available online in very many languages. Although these are translated texts, the brochures are meant to convince people, so the translations should be made such as to feel natural to the readers. Still, there will be influences from translationese in these texts, so they are not suitable for the investigation of the details of role marking in individual languages. However, for the purpose of comparing languages these texts are highly suitable, because they present a clearly comparable resource across languages.

For this paper, I will restrict myself to bound case-like marking only, likewise purely because of practical reasons. So, languages without bound case-like marking are uninformative for this paper. I deliberately use the term ‘case-like’ marking, because I define such marking pragmatically for this study on a purely orthographic basis. Whatever is written as one word together with a nominally used root is included here as ‘case-like’ marking. For future research, a more linguistically adequate and more all-encompassing notion of flagging and cross-referencing of noun phrases should be considered. However, even with this limited notion of linguistic marking, it turns out that there is still enough information to induce various semantic roles. In general, it seems to be the case the rather coarse-grained linguistic notions are already sufficient to investigate the typological diversity of the world’s languages, though it should be realized that such rough approximation of linguistic structure are of course not suitable for the study of the structure of individual languages.

To easily find comparable roles across the various translations, I have investigated the marked forms of the word ‘bible’. First, this word occurs with a high frequency in the religious brochures from watchtower.org, so sufficient data can already be found in a rather short text. A further profitable aspect of using the word ‘bible’ is that the bible takes on a great variety of roles in the pamphlets. The bible occurs in agent-like roles, as in “the bible teaches us”, but also in undergoer-like roles, as in “you should study the bible”, or in various other roles, as in “the bible’s view” or in “to have respect for the bible”. This variety of roles offers a suitable background for the investigation of variation in role marking across the world’s languages. Finally, because of its high frequency and its often rather obvious form, the word ‘bible’ is easily recognizable, also in languages which I am not able to read myself.

In practice, I selected 34 contexts in the pamphlets in which the word bible occurs. Various possible contexts were removed from the selection because the actual word for ‘bible’ was not used in a sufficient number of languages (only cross-referencing was used in some contexts in some languages). The English and German translations of the chosen 34 contexts are shown in Appendix A.

Shown in Table 1 are the 15 languages sampled for the current paper. A map of the geographic locations of these languages is shown in Appendix B. The information on genealogical affiliation (genus, family), geographic location and macroarea are taken from the *World Atlas of Language Structures* (WALS, (Haspelmath *et al.* 2005)). The languages show a wide variety of alignment structures, as summarized in the last column of the table. The alignment of Oromo, Khoekhoe, Irish, Korean, Drehu, Nias, Greenlandic, Aymara are due to (Comrie 2005). Albanian, Faroese, Estonian, Azerbaijani are relatively straightforward nominative-accusative language like all Indo-European languages in Europe. The ergative alignment of Akha is discussed in Terrel (2009). Ma'di is normally not considered to have case marking (Crazzolara 1960: 20), and the nominal suffix *-i* which caused the inclusion of this language in the sample is commonly analyzed to be some kind of focus marking. Likewise, the Irish initial consonant mutation (which is the bound marking attested in the world for 'bible') is normally not considered to be role marking, but it's behavior will turn out to be very similar to 'regular' case markers of other Indo-European.

Language	Genus	Family	Macroarea	Alignment
Oromo	Eastern Cushitic	Afro-Asiatic	Africa	Marked nominative
Khoekhoe	Central Khoisan	Khoisan	Africa	Nominative-accusative
Ma'di	Moru-Ma'di	Nilo-Saharan	Africa	Neutral
Albanian	Albanian	Indo-European	Europe	Nominative-accusative
Irish	Celtic	Indo-European	Europe	Neutral
Faroese	Germanic	Indo-European	Europe	Nominative-accusative
Estonian	Finnic	Uralic	Europe	Nominative-accusative
Altai	Turkic	Altaic	Asia	Nominative-accusative
Azerbaijani	Turkic	Altaic	Asia	Nominative-accusative
Korean	Korean	Korean	Asia	Nominative-accusative
Akha	Burmese-Lolo	Sino-Tibetan	Asia	Ergative-absolutive
Drehu	Oceanic	Austronesian	Pacific	Active-inactive
Nias	Sundic	Austronesian	Pacific	Marked absolutive
Greenlandic	Eskimo-Aleut	Eskimo-Aleut	America	Ergative-absolutive
Aymara	Aymaran	Aymaran	America	Marked Nominative

Table 1. Language sample for the current study.

4. Analysis of roles

The actual wordforms as attested for the word ‘bible’ in the current language sample are shown in Appendix C. This appendix represents the basic data for the further analyses to be performed in this paper. There are various calculations that can be performed based on the distribution of different forms across the contexts.

First, the marking of contextual roles can be compared by investigating their language-specific encoding. By simply counting how often two contextual roles are marked differently in the languages sampled (and dividing this by the number of comparisons made) an average role similarity can be established (cf. the isomorphism hypothesis from Section 2). For example, between the first and the second contextual role of my selection, there are 10 language that use a different form, so the average distance between these two contextual roles is 10/15 = 0.67. These computations are performed for all pairs of context, and the resulting distances are shown in Appendix D. Note that for the establishment of these distances, there has not been made any typological comparison *between* the languages. Only forms *within* each language have been compared to each other. There was no decision necessary which forms from language A should be compared to which forms from language B.

This distance matrix between the contextual roles represents a semantic map on these roles, though without a graphical representation yet (for a more detailed explanation why this really is a semantic map, see Cysouw 2010). There is a multitude of possibilities to graphically represent the distance matrix. Figure 1 shows the first two dimensions of a multidimensional scaling (MDS) of the distance matrix. The position of the numbers in the figure is determined by the MDS, showing similar predicate-specific roles as being close to each other. The circles and the annotations in this figure have been added manually to indicate the close approximation of the statistical analysis to the predicate-based notion of roles. Remember that at no point in the analysis any information about the lexical verbs was used to determine the positioning of the points in the figure.

These first two dimensions of the MDS as shown in Figure 1 are actually still not a particularly good approximation of the variation, as they only represents about 50% of the eigenvalues. Other possibilities would be a hierarchical clustering scheme like NeighborJoining (Saitou & Nei 1987) or a split decomposition (Bandelt & Dress 1992) like NeighborNet (Bryant & Moulton 2004). Such pictures are shown in Appendix E. For the purpose of this paper, I will only use the MDS display, as it allows to overlay other information on top of the two dimensional representation.

Another way to analyze the distance matrix of the contextual roles is to perform 'flat' clustering, i.e. divide the contextual roles into mutually exclusive groups of similar roles. For such an analysis, one has to pre-set the number of clusters, and then an optimal division of the contextual roles into those clusters is determined. However, not every number of clusters is equally adequate. The suitability of a flat clustering is determined, roughly speaking, by strong internal similarity within each cluster and clear separability between the clusters. I will use here the 'partitioning around mediods' (PAM) clustering approach by Kaufmann & Rousseeuw (1990), with the associated measure of suitability of the clustering (the 'average silhouette width'). The suitability of clustering for all number of clusters from two to 30 are shown in Figure 2. The optimal clustering is found with ten clusters, while there are suboptimal maxima at seven and three clusters.

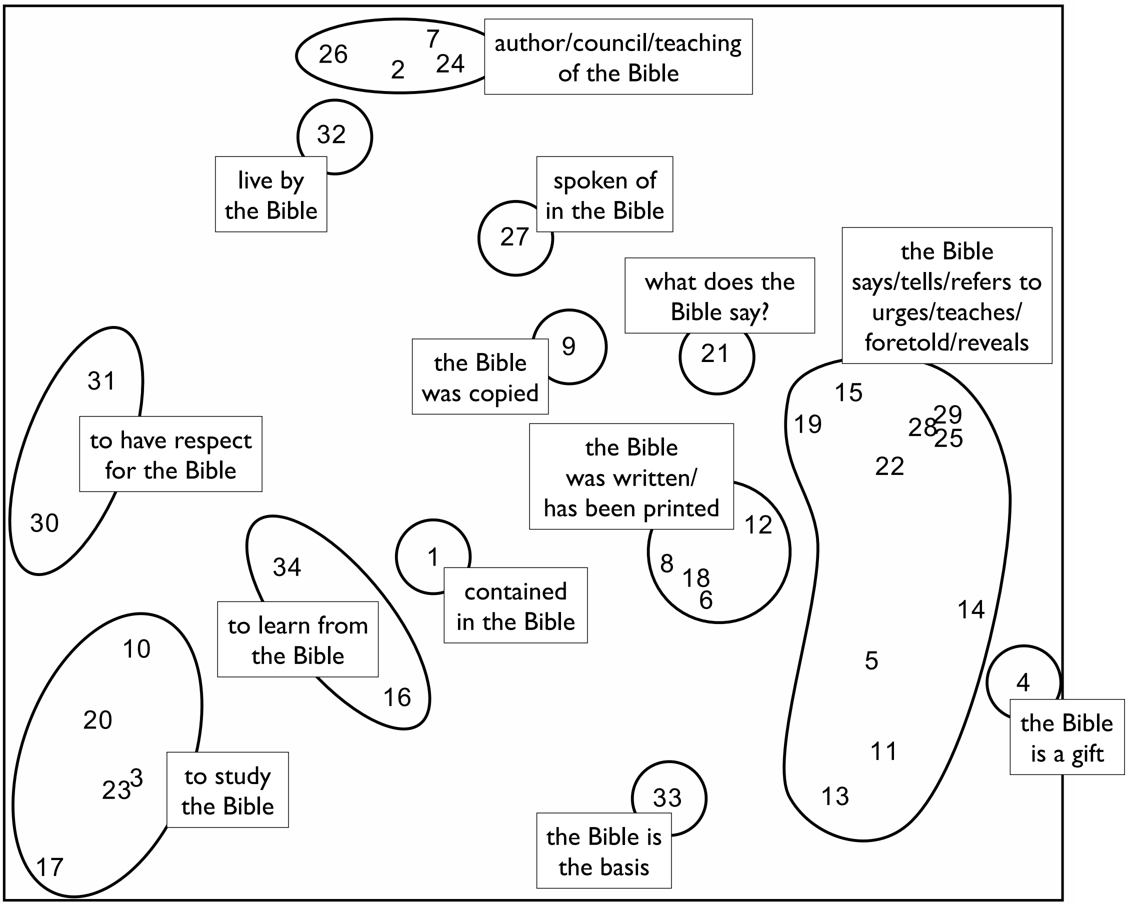

Figure 1. Semantic map of contextual roles, with hand-drawn clusters of approximate lexically-specified roles.

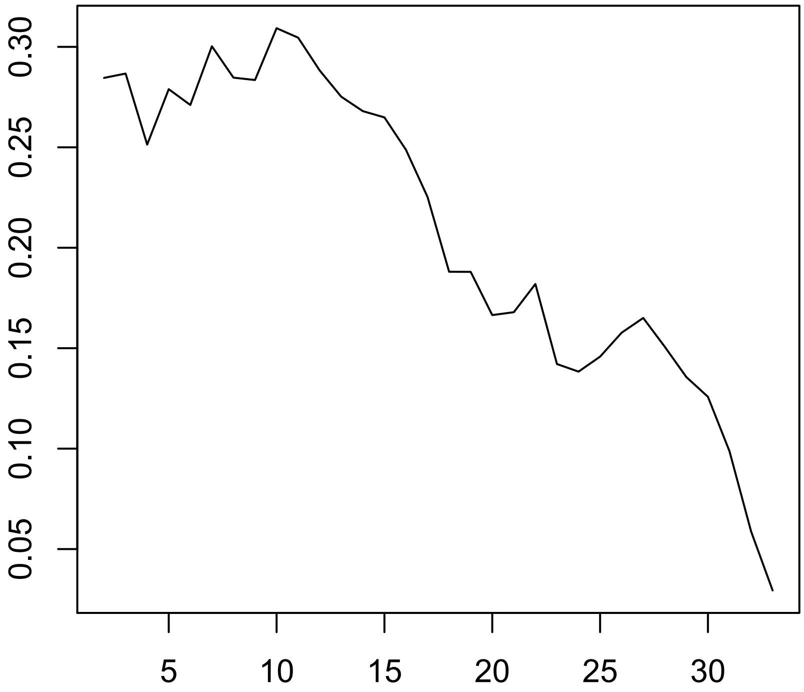

Figure 2. Suitability of the optimal clustering for different number of clusters. Shown on the x-axis is the number of clusters, while the y-axis shows the suitability of this ‘flat’ clustering in the form of the ‘average silhouette width’. The best clustering is attested with 10 clusters, while 7 and 3 clusters are other good choices.

The clustering of all 34 contextual roles into three groups as suggested by the PAM-method is shown in Figure 3 (the clustering into ten and seven groups are not shown here for reasons of space, and can be found in Appendix F). This figure uses the same MDS display of the 34 roles as used in Figure 1, only the superimposed groups are different. This attested clustering shows a striking parallel to the intuitive notion of macro-roles. Note that because the MDS and the clustering are different mathematical methods that focus on slightly different numerical aspects of the underlying data, the visual impression as shown in Figure 3 looks slightly inconsistent, especially concerning the placement of contextual role number one. However, this simply represents a role with rather undetermined correspondence to other roles, which results in a placement in the middle of the MDS. The English translation of the sentence in which this role occurs is “What important information is contained in the Bible?”, which is in many languages translated without the passive construction as found in English (e.g. German “Welchen wichtigen aufschluß enthält die Bibel?”).

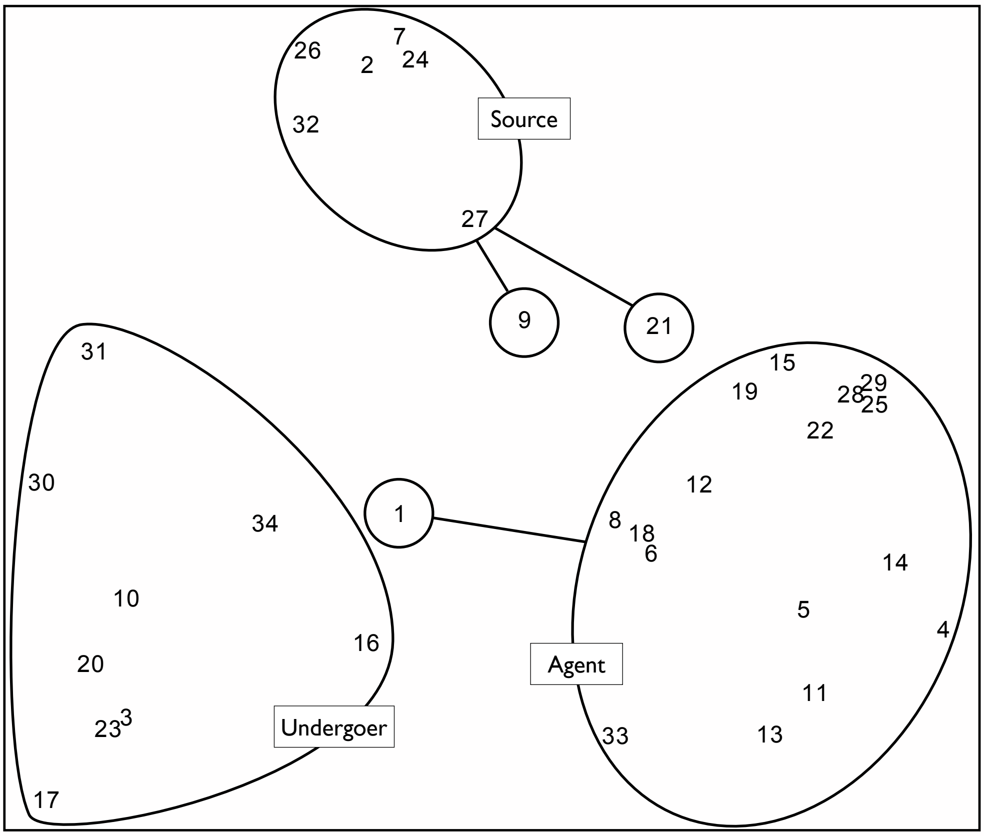

Figure 3. Clustering of the contextual roles into three clusters, which strongly correlate with cross-linguistic macro-roles Agent, Undergoer and Source. The clusters are depicted on the same MDS basis as Figure 1.

5. Comparison of languages

As can be seen in the language-summary as presented in Table 1, there is a wide variety of alignment patterns (of full noun phrase marking) available in the fifteen sampled languages. The largest group has nominative-accusative alignment (seven languages). There are also two languages with ‘marked’ nominative-accusative alignment, which are languages in which—unexpectedly from a typological perspective—the patient roles are formally more marked than the agent roles (Handschuh 2011). Further, there are three languages with ergative-absolutive alignment, among which there is a single ‘marked’ ergative-abolutive one. Finally, there is one language that is analyzed as active-inactive, and two language that are normally not analyzed as being case-marked, and thus are of neutral alignment.

Although such typological distinctions suggest strict categorical differences between the languages, the attested differences are mostly much more continuous in nature. Traditional typology relies heavily on a small set of strictly selected indicators for the establishment of types. Specific characteristics are selected to define types and to classify languages accordingly. Further, the classificatory decision are mostly made on the basis of secondary sources (i.e. descriptions of the languages in question), and not on the basis of actual comparable examples.

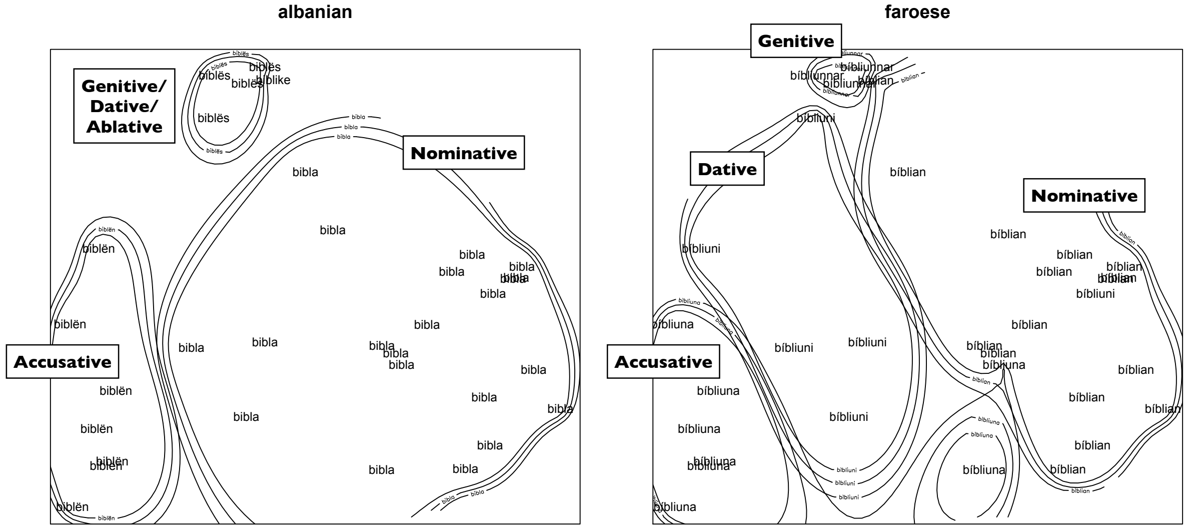

Figure 4. Language-specific coding of the contextual roles, illustrated for Albanian and Faroese. The position of the forms is identical to the numbers in Figure 1. The clusters are drawn using an interpolation technique called ‘Kriging’. The labels are language-specific labels as used for the description of these languages.

The current dataset offers the possibility to perform a much more detailed typological comparison. To understand how it is possible to make such comparisons of complete languages, consider the semantic maps of Albanian and Faroese, as shown in Figure 4. These figures use the same MDS layout of the 34 contextual roles as was also used in earlier figures. However, instead of plotting the numbers referring to the clauses, in these figures the actual case-marked forms as attested in the text are shown. To show the language-specific structure of this coding, I have added automatically drawn clusters around identical forms, resulting in a special kind of semantic maps. These clusters were established by first making a 3D interpolation for each case-marked form, in which the height of the interpolation is established by the density of the occurrence of the form in the MDS base-map. Basically, the more forms occur close to each other, the higher the ‘mountain’ will become. This mountain is then drawn in the form of height lines at heights 0.45, 0.50 and 0.55 (which results in the slightly fuzzy appearance of the borders). More details about this approach to draw semantic maps can be found in Cysouw & Forker (2009). Labels were manually added to identify the clusters. Note that these labels have are capitalized as they are names for language-specific structures and not cross-linguistic categories. I have produced such semantic maps for all languages in the sample, using exactly the same graphical settings so the resulting pictures can be visually compared to each other. For reasons of space and because such semantic maps are much easier to interpret when using colors, the pictures are not included in this printed article, but can be found in Appendix G.

Looking at Figure 4, the case marking structure of the two languages seems pretty much alike, pace for the addition of a separate Dative in Faroese. This impression of relative similarity between two languages can be easily formalized into a general measurement of language similarity. Basically, for each language I consider all 1122 (= 34 × 33) pairs of contextual roles, which can either have identical (= 1) or different (= 0) case marking (see Cysouw 2010 for more details on the establishment of such language-specific metrics). Two languages can be compared by comparing these 1122 pairs between the two languages, e.g. by taking a Pearson correlation coefficient between them. This similarity between two languages can then be computed for all pairs of languages (see the results in Appendix H), and the resulting metric on the languages can be interpreted as a ‘typology without types’. In such a typology, all languages are compared to each other, and the resulting grouping of languages can be investigated with various statistics techniques, just as already roughly outlined in Section 4. Shown here in Figure 5 is a NeighborNet illustrating the structure of the similarities between the languages.

There are various interesting observations to be made on the basis of this ‘typology without types’. First, the languages to the left include two ergative languages (Nias, Akha), but also Ma’di and Drehu, which are not normally considered to be ergative. Looking at the semantic maps for these languages (cf. Appendix G), the characteristic binding these languages together is the existence of a case-like marker that is used in a wide variety of contextual roles (spanning almost the complete set of 34 contextual roles sampled), including all typically patient-like roles (cf. Figure 3). This can be interpreted as that these ‘ergative’ markers are functionally unmarked (i.e. they occur in most contextual roles). This makes also sense for the analysis of Ma'di, as the case-like markers in question here are traditionally analyzed as being markers of information structure. However, the unmarked form occurs in a wide variety of context, while the marked 'focus-marking' suffix mainly occurs in transitive agent like contexts.

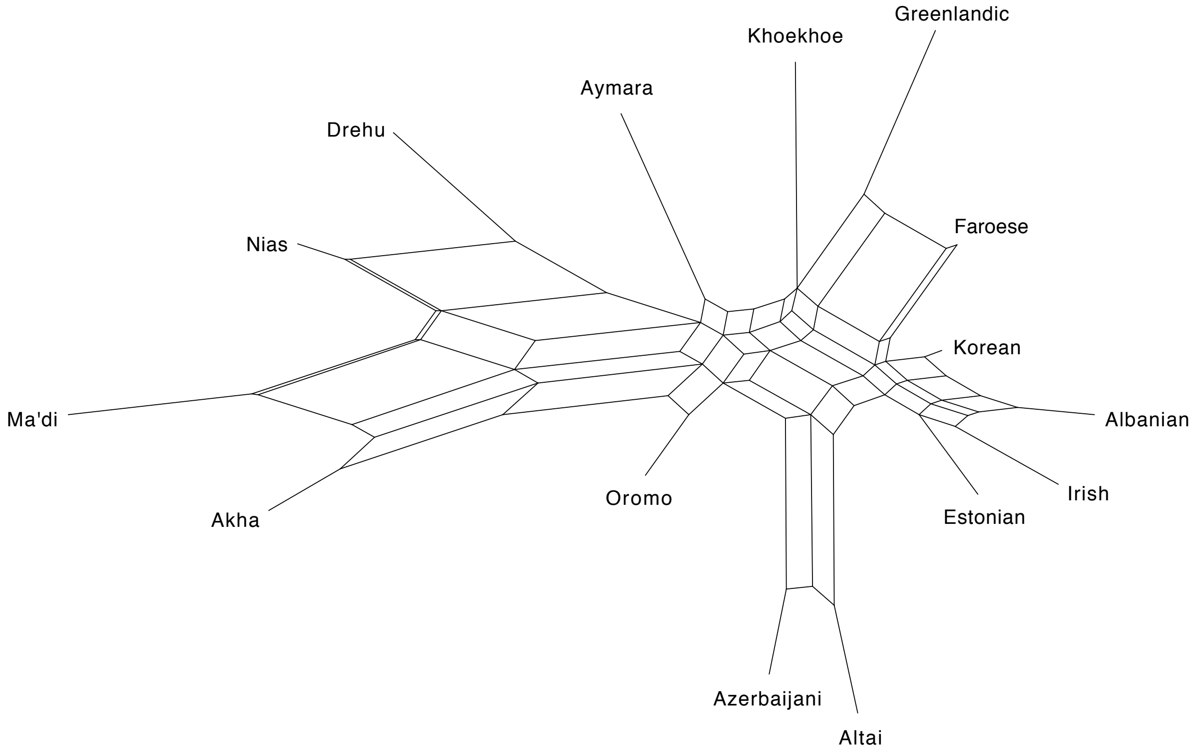

Figure 5. NeighborNet of languages according to their similarity in case marking structure. The languages to the left typically have a widespread marker that is also used for patient-like contexts. This group includes typically ergative languages. In contrast, the languages to the right are typically nominative languages, while Aymara and Oromo are of the 'marked' nominative type. Greenlandic is unexpectedly grouped with the nominative languages.

Further note that Greenlandic, which is traditionally defined as being ergative, does not occur close to these languages on the left side. This is basically due to the fact that the structure of marking is completely different in Greenlandic. Although there is a specific case also used for transitive agents (the traditional definition of ergativity), there is no complementary 'unmarked' case used for a large group of the remaining roles.

All languages on the right are basically nominative-accusative languages, though Azerbaijani and Altai seem to stand out. It is not clear to me why this happens, as the structure of the case marking in these languages does not very much look the same. However, they both seem to have rather different structures from all other languages, so they might have been grouped together because of their shared dissimilarity from all other languages considered here.

Finally, Aymara and Oromo are found in between the ergative languages to the left and the nominative languages to the right. Both these languages are considered to be ‘marked’ nominative in that formally the marking of the nominative is overt, while the accusative is formally unmarked. For Oromo this formal marking structure is also reflected in the functional marking structure, as the unmarked patient-like case marker also shows a wide distribution over the 34 contextual roles. The reason for the intermediate status of Aymara is not immediately obvious to me.

In summary, it is possible to statistically classify languages on the basis of their language-specific marking of contextual roles. The resulting alignment typology is somewhat alike to the traditional nominative-ergative typology, though much more emphasis is put on the extend of the distribution of the cases. Language with the same kind of distribution of cases over roles are grouped together, which in practice gives a stronger weight to similarity between forms that are widespread (i.e. functionally unmarked) and not assigns much influence to details of the highly specific marked structures.

6. Conclusion

Based on an admittedly rather limited data set, this paper has shown the viability of using contextual roles as a basis for the typological comparison of roles in the world’s languages. Contextual roles are the actual roles as they occur in context. Such roles are of course strongly determined by the lexical predicate used in the sentence, but also other linguistic coding implicitly is included in the determination of the marking. To be able to compare contextual roles across languages it is necessary to have access to some kind of parallel text, be it in the form of translations (as used in the present study) or in the form of more experimentally controlled parallel utterances (e.g. using pictures, films, or other stimuli).

Clustering of the formal marking of these parallel contextual roles offers the possibility to statistically derive higher-level role abstractions, very much alike to traditional predicate-based roles or even macro-roles. From the same data it is also possible to establish similarities between complete languages, resulting in a ‘typology without types’, i.e. a measurement of fine-grained similarities between languages from which more coarse-grained typological clusters of languages (alike to traditional ‘types’) can be derived. Looking forward, the fine-grained typology does seem to offer fascinating possibilities to much more accurately capture the real diversity of the world’s languages, which normally only under protest agree to be classified into a few broad all-encompassing types. The real challenge for future research is not to formulate such fine-grained typologies, but to successfully show how to they can elucidate correlations and/or restrictions on linguistic structures.

Appendices

The following appendices with all data and other supplemental material can be accessed online at the Open Data Repository of the LMU München at http://data.ub.uni-muenchen.de/.

Appendix A: Sampled contexts.txt
Appendix B: Map of languages.pdf
Appendix C: Wordforms.txt
Appendix D: Contextual role distances.txt
Appendix E: Clustering of contextual roles
Appendix F: Alternative flat clustering
Appendix G: Language specific structures
Appendix H: Language distance.txt

References

Bandelt, Hans-Jorgen & Andreas W M Dress. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1(3). 242-252.

Blake, Frank R. 1930. A semantic analysis of case. *Language* 6(4). 34-49.

Bryant, David & Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255-265.

Comrie, Bernard. 2005. Alignment of case marking of full noun phrases. In: Martin Haspelmath, Matthew S Dryer, David Gil & Bernard Comrie (eds.), *World Atlas of*

Language Structures, 398-405. Oxford University Press: Oxford.

Crazzolara, J P. 1960. *A Study of the Logbara (Ma'di) Language.* Oxford University Press: London.

Cysouw, Michael. 2010. Semantic maps as metrics on meaning. *Linguistic Discovery* 8(1). 70-95.

Cysouw, Michael & Diana Forker. 2009. Reconstruction of morphosyntactic function: Non-spatial usage of spatial case marking in Tsezic. *Language* 85(3). 588-617.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67(3). 547-619.

Fillmore, Charles J. 1968. The case for case. In: Emmon Bach & R T Harms (eds.), *Universals in Linguistic Theory,* 1-88. Holt, Rinehart and Winston: New York.

Handschuh, Corinna. 2011. *A typology of marked-S languages.* University of Leipzig: Leipzig. Ph.D. Thesis.

Kaufman, Leonard & Peter J Rousseeuw.1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* (Series in Applied Probability and Statistics). Wiley: Hoboken, N.J.

Martin Haspelmath, Matthew S Dryer, Bernard Comrie & David Gil (eds.).2005. *The World Atlas of Language Structures,* Oxford University Press: Oxford.

Saitou, Naruya & Masatoshi Nei. 1987. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4). 406-425.

Terrell, Jake. 2009. Semantic case marking in Akha. *University of Hawai'i Department of Linguistics Working Papers in Linguistics* 40(3). 1-11.

Van Valin, Robert D Jr. 2004. Semantic macroroles in Role and Reference grammar. In: Rolf Kailuweit & Martin Hummel (eds.), *Semantische Rollen,* 62-82. Narr: Tübingen.

Wälchli, Bernhard & Michael Cysouw. 2011. Toward a semantic map of motion verbs. *Linguistics* (forthcoming).