Capturing diversity in language acquisition research

Sabine Stoll & Balthasar Bickel
University of Zürich

Abstract
In order to understand how children cope with the enormous variation in structures worldwide, developmental paths need to be studied in a sufficiently varied sample of languages. Because each study requires very large and expensive longitudinal corpora (about one million words, five to seven years of development), the relevant sample must be chosen strategically. We propose to base the choice on the results of a clustering algorithm (fuzzy clustering) applied to typological databases. The algorithm establishes a sample that maximizes the typological differences between languages. As a case study, we apply the algorithm to a dozen typological variables known to have an impact on acquisition, concerning such issues as the presence and nature of agreement and case marking, word order, degrees of synthesis, polyexponence and inflectional compactness of categories, syncretism, the existence of inflectional classes etc. The results allow deriving small samples that are maximally diverse. As a side result, we also note that while the clustering algorithm allows maximization of diversity for sampling purposes, the resulting clusters themselves are far from being discrete and therefore do not reflect a natural partition into basic language types.

1. Introduction

After many decades in which universals have played the main role in both typology and language acquisition research, a hitherto much neglected feature of human languages has taken center stage again: diversity. In typology, claims about universal grammar have been increasingly disputed and, elaborating on this trend, Evans & Levinson (2009) argue that cognitive scientists have not been sufficiently aware of the range of linguistic diversity and therefore have artificially limited the domain of human cognitive abilities that they can possibly cover. Evans and Levinson argue that languages can vary in many more and many more substantial dimensions than commonly assumed in the cognitive sciences and that this defines new and important challenges: How can structures with this range of variability be parsed and

learned, given the genetic unity of our species? From this perspective, one of the most pressing tasks in language acquisition research is to explain how children can cope with the variation in the languages of the world.

This contrasts with what has dominated language acquisition research for a long time: the assumption of a genetically given Language Acquisition Device (LAD), containing the abstract syntactic principles that are preconditions for the acquisition of any conceivable human language (Chomsky 1968, 1980). The assumed content of the LAD has changed drastically over the years. In earlier versions, the LAD was assumed to consist of formal principles organizing grammar, while later developments added parametric variation (Chomsky 1981; Hyams 1986; Borer & Wexler 1987). In current proposals, the LAD is often narrowed down to one single feature, namely recursion (Fitch et al. 2005). However, even this last universal has not remained unchallenged, and it is an open issue whether recursion is specifically linguistic or even specifically human (Everett 2005, 2009; Evans & Levinson 2009; Van Valin 2009). Interestingly, the reduction of the LAD to the single issue of recursion narrows the gap between generativist and non-generativist approaches to language acquisition: for both approaches, the main challenge now is to explain how children can cope with the known variation of structures, and for all practical purposes, it seems a relatively minor issue whether recursion is part of the variation or whether recursion is the one mechanism that children don't have to learn from the input.

In this paper, we propose a new strategy of sampling languages that allows more systematic attention to cross-linguistic variation in language acquisition. We first discuss the need for this attention in more detail (Section 2) and briefly summarize the state of the art in cross-linguistic acquisition research (Section 3). After discussing traditional sampling methods in typology in Section 4, we introduce our own method in Section 5 and illustrate it by way of a case study targeted to research on complexity in acquisition in Section 6. We then discuss further interpretations of the results in Section 7 and draw general conclusions in Section 8.

2. Variation and language acquisition

In order to make any progress in studying language acquisition as a general human feature, we need a full appreciation of variation in languages and, therefore, need acquisitional data from languages as diverse as possible. The sample of languages is important because the structure of the languages under investigation has a far-reaching impact on theory. In the past decades, research has typically focused on the acquisition of some regular features of English grammar and this influenced theory building to a significant degree. The focus on English grammar resulted in the theoretical claim about language learning as a rule-based mechanism, and this has been advocated by

many researchers as the predominant general linguistic learning mechanism (most pronouncedly by Pinker 1994).

Rule-based learning seems to work well for instance in describing the acquisition of the English plural (Berko Gleason 1958) or the English past tense (Pinker & Ullman 2002). The English plural is structured in such a way that there is a large number of regular forms for which rule-based learning can be postulated, and only a small number of irregular forms which have been claimed to be learned like lexical items. The postulation of a general rule-based learning mechanism for the acquisition of the plural and the distinction between regular and irregular forms, however, is most likely an artifact of English grammar. The theory does not generalize to even closely related languages, such as German. Studies on the acquisition of the German plural suggest that German plural acquisition is better explained by a schema model, and only a systematic comparison of English and German was able to show that schema learning can explain the acquisition paths of both languages (Köpcke 1998; Behrens 2002). The distinction between regular and irregular forms and thereby the role of rule-based learning more generally becomes even more questionable in languages that are very different from the typical Indo-European languages, such as, for instance, Dinka (Western Nilotic) in which we do not find distinctions between regular and irregular number morphology at all (Ladd et al. 2009).

The assumption of rule-based learning is also difficult to maintain in other grammatical domains, such as for instance in the acquisition of the grammatical category of Russian aspect. It is very well possible to postulate several rules for determining the aspectual values of Russian verb forms, but it is far from obvious how many regular forms should be postulated and which rules children would in fact derive from the data when learning the relevant aspectual distinction. In this case again, rule-based learning does not explain well the acquisition process, and a context-based approach performs better (Stoll 1998, 2005). These examples show that variation is a crucial testing ground for theories, and theories that have been tested only on one language underestimate the learning tasks involved in language learning more generally.

Lately, cross-linguistic studies of language acquisition on a large variety of languages have indeed shown evidence that typological variables influence the course of acquisition, and to a certain extent, the conceptualization of reality. Earlier assumptions of universals or of learning strategies that are based on the primacy of a putatively universal conceptual development have found little support. Rather, language-specific factors seem to be more relevant for the acquisition process. Even those cognitive domains that were traditionally assumed to be uniform across human societies, such as spatial cognition, have recently been shown to exhibit an impressive amount of variation across cultures and languages (Bowerman & Levinson 2001; Levinson 2003). Depending on how a language structures space, children focus on the

relevant features relatively early in their development (e.g., Choi & Bowerman 1991). Further, in their earliest development children have been noted to express motion events according to the patterns found in their native language and not according to what are sometimes assumed to be universally uniform conceptual patterns (Berman & Slobin 1994).

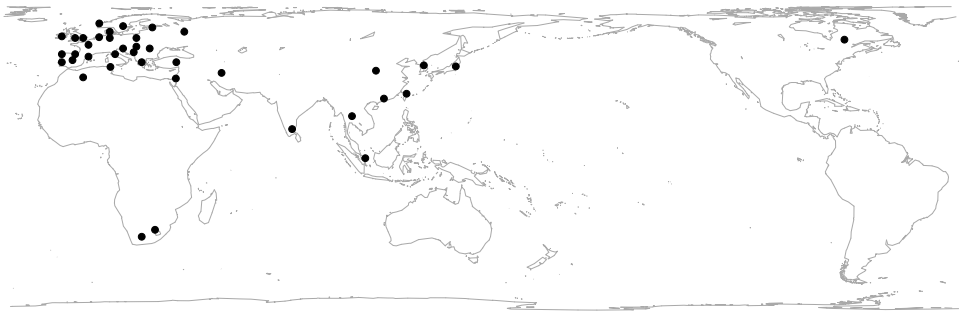
Another domain for which universal behavior has been assumed but that has been challenged lately is the preference for nouns in early acquisition. It has been claimed (Gentner 1982; Gentner & Boroditsky 2001) that nouns are learned more easily and earlier than verbs because things are assumed to be conceptualized in easier ways than events and that this represents a cognitive universal. Recent research, however, has shown that here is no universal preference in whether children focus in their early vocabulary on nouns or verbs. What matters instead are language-specific issues such as the distribution of nouns and verbs in the input (Tardif 1996), differences in the morphological complexity of these categories (Stoll et al. 2012) and language-teaching strategies (Choi 2000; Tardif et al. 1997). These findings again show that linguistic variation has a key role to play when we try to understand how children acquire language.

3. The data scarcity problem

However, the main problem in studying cross-linguistic diversity in acquisition research is the scarcity of data we have. Despite efforts by Dan Slobin and colleagues to increase our database on language acquisition over the past three decades, the sample of languages for which we have language acquisition corpora is still very limited and makes up less than 0.1% of the 6000-7000 languages spoken today (Stoll 2009; Lieven & Stoll 2010). There are corpora from 37 languages publicly available.[1] These corpora are transcribed. Some of them are morphologically glossed, and only very few come with translations into better known languages which complicates comparative research. There exist several privately owned corpora, which are, however, not readily accessible.

What is worse is that the available corpora are very much biased toward Indo-European languages spoken in Europe (Map 1). This bias has not only an effect on the research questions we can ask but also on the results we get and the subsequent generalizations we make. Some of the very prominent features of European languages are quite rare outside of Europe, such as a definite versus indefinite article distinction, relative pronouns, 'have'-based perfects, participle-based passives, and so on (van der Auwera 1998; Haspelmath 1998, 2001), but it is exactly about the acquisition of these features that we know a lot about, at the expense of other, more widespread features in the rest of the world. In addition, the languages spoken in Europe are

[1] Via CHILDES (Child Language Data Exchange System), `http://childes.psy.cmu.edu/`.



Map 1. Languages for which there are longitudinal corpora in the CHILDES database (as of 2013).

quite homogenous in many respects. As a result, in concentrating predominantly on features of European languages, we get very biased results and miss the full range of diversity of acquisition challenges. Since a great percentage of the languages spoken in the world are endangered, it is high time to document not only the grammar but also the acquisition process. If we aim at comparative work using longitudinal corpora, however, the choice of languages is complicated. The development of a corpus is too expensive an enterprise to be limited to the study of a single or a small set of variables as one would do in an experiment. Instead, a corpus is expected to support research on many different aspects of acquisition. This makes it unclear which languages one should choose when setting up a comparative sample of corpora. As a result, acquisition research has to address similar sampling issues already well known in typological research.

4. Traditional approaches to sampling

When typology started to adopt quantitative methods, sampling became an increasingly important issue (Bell 1978). Given the fact that the language documentations available are not distributed evenly across language families and that the documentation of these languages also varies substantially in quantity and quality, a number of biases are virtually unavoidable: bibliographical bias (strong bias away from smaller languages and families and areas that are difficult to access), genealogical bias (as a result of the bibliographic bias), areal bias, typological bias and cultural bias (Rijkhoff & Bakker 1998; Bakker 2011). The available sample of languages we have can thus hardly achieve representativity.

Rijkhoff et al. (1993) identified two main approaches to sampling in typology: one approach aims at statistical estimates on frequencies, preferences, or areal patterns in the languages of the world. This type of sampling is called probability sampling. Various sampling methods of this type are discussed and illustrated in Bell (1978), Perkins (1989), Dryer (1989), Bickel

(2008), among others. The second approach aims to capture the extent of variations, regardless of probabilistic estimates of feature distributions. Rijkhoff and colleagues call this approach diversity sampling, and it is this type of approach we focus on in this paper.

The diversity sampling method proposed by Rijkhoff et al. (1993) is based on genealogical taxonomies. The internal structure of the taxonomies is used to estimate diversity within families. Depending on the depth and the width of the tree, the number of languages for a given family is taken to define the family's proportion in the required sample. A key problem of this approach, however, is that genealogical taxonomies are not necessarily a reliable indicator of typological diversity: as Nichols (1996) has emphasized, the kind of variables that define genealogical groups and tree shapes have a very different nature from the kind of variables that define typological diversity. For genealogical grouping one relies on arbitrary, idiosyncratic features that (ideally) meet Nichols's statistical threshold for an individual-identifying feature (also see LaPolla, this volume). By contrast, typological variables are typically not idiosyncratic, often externally (functionally) motivated, and therefore orthogonal to the genealogical taxonomies.

In the following, we develop an alternative method for diversity sampling, specifically tailored to the needs of comparative language acquisition research but with possible applications beyond this. Rather than relying on genealogical taxonomies, our method is based on typological variables.

5. An alternative: Clustered Sampling

The key challenge in sampling languages for developing longitudinal corpora comes from three conflicting constraints:

1. The sample size should be as small as possible because the development of longitudinal corpora is extremely time consuming and expensive.

2. The sample size should be large enough to cover as much structural diversity as possible.

3. The sample should allow contrastive studies of pairs (or other small groups) of languages in which many variables can be kept constant and only few variables vary.

The first concern is a practical one, but it severely limits what can realistically be done. In our own experience, developing a longitudinal corpus of a single, previously under researched language takes five to seven years, a dozen native speaker assistants for transcription and translation work, and another dozen linguists for grammatical tagging. The reason for this is that longitudinal corpora are useless for acquisition research unless they cover a sufficiently long period (at least one year) and that for statistical purposes

the individual recording sessions during the period need to be fairly long (at least an hour per month). Further, corpora should not be limited to a single child because there is substantial variation between individuals, and there is always a nonnegligible chance that a child's development is atypical.

The second constraint derives from the theoretical desideratum that a full understanding of language acquisition must include an account of how children learn *any* kind of structure, as common or uncommon it may happen to be in the languages of the world. In other words, ideally we need full coverage of diversity. The sample sizes that this requires are in the thousands, which is obviously in conflict with the first constraint.

The third constraint is motivated by the kinds of questions that are of interest to comparative language acquisition research: in order to determine the impact of a specific variable on acquisitional pathways, we need to be able to keep other variables as constant as possible. For example, in order to find out the impact of differences in verb morphology, we need to keep variables of word order constant because there is evidence that verb morphology is learned differently in verb-peripheral versus verb-medial languages (see Stoll et al. 2012, for discussion). The most straightforward and most economical way of making such research possible is by forming pairs of languages that differ in one variable but are identical or at least very similar in many others. This again leads to large sample sizes: already when considering ten variables, allowing three types (values) each (e.g. verb-initial, verb-medial, and verb-final order; or separative, cumulative, and distributive exponence of tense markers), there are thirty possible pairs of languages such that each pair differs in only one of the ten variables.[2]

In order to find an optimal compromise between the three constraints, we translate the sampling problem into a problem of cluster analysis because cluster analysis shares our general goal: grouping data into sets such that the variation within sets is smallest (Constraint 3) and variation between sets is largest (Constraint 2). Constraint 1 defines the upper limit of sets, and this corresponds to the special family of cluster analysis known as partitioning methods (Kaufman & Rousseeuw 1990): these methods start from a predefined number of sets and allocate the data to these sets so as to minimize within-set and maximize between-set variation.

For our purposes, however, we are not so much interested in the exact boundaries of the sets. What is more interesting is to find sets with prototype effects (of the kind familiar from categorization research): we can find an optimal balance between the three constraints if we can partition languages into few sets (Constraint 1) in such a way that each set contains two or three prototypes (best exemplars) that are very different from the prototypes of other sets (Constraint 2) but very similar among themselves (Constraint 3).

[2] Generally, n variables $V_i \dots V_n$ with k_i types allow $\sum_{i=1}^{n} \binom{k_i}{2}$ such pairs, which grows factorially with k and additively with n.

Fortunately, there is a well-established type of partitioning method with exactly these properties: fuzzy clustering. The method is based on the formal notion of fuzzy sets, that is, sets to which members belong with a certain probability between 0 and 1 (the membership coefficient). Kaufman & Rousseeuw (1990) introduced an iterative algorithm for optimizing membership coefficients such that they minimize within-set dissimilarity (distance) and maximize between-set dissimilarity, given a predefined number of sets (also cf. Maechler et al. 2005). Dissimilarities can be estimated by distance measures such as Hamming distances, that is, by the proportion of variables in which languages have the same values. The result is a list of sets and estimates of the degree to which each language belongs to any one of these sets, for example, language L_1 may belong to set A to 85% and to B to 15%; L_2 to A to 60% and to B to 40%; L_3 to A to 10% but to 90% to B and so forth.

From each set, we can then extract the two languages that have the highest membership coefficients (but possibly also considering practical issues like degrees of endangerment): these are the prototypes of the sets and will necessarily be very similar to each other (satisfying our Constraint 3). By virtue of best representing their sets, they at the same are maximally different from the prototypes of all other sets (satisfying Constraint 2).

In the following section, we illustrate the method through a case study that is of key interest to acquisition research: variation in complexity.

6. An example: a clustered sample capturing complexity variation

As Nichols (2008) has proposed, complexity chiefly arises from the number of variants available to language users and from the amount of information needed for explaining the choice among variants. From an acquisitional point of view, the relevant areas of grammar are chiefly morphology and word order: complexity in morphology is known to slow acquisitional curves, but this also interacts with the position of morphology-rich forms at the periphery, as opposed to the middle of utterances because periphery positions are known to be more salient psychologically (Slobin 1973, 1985; Tardif et al. 1997; Stoll et al. 2012:among others). In response to this, we selected for our case study a series of morphological variables and one word order variable:[3]

VERB POSITION: A four-way distinction between verb-initial, verb-medial, verb-final, and unconstrained structures as the basic (most common, default) word order of a language. *Data source:* Dryer (2005b) and an expanded dataset from Nichols (1992) in AUTOTYP.[4]

[3] Where data come from different databases, we merged them if the proportion of languages for which the databases had conflicting information was less than 5% of the number of languages for which the databases had shared information at all.

[4] See `http://www.wals.info` and `http://www.uzh.ch/spw/autotyp`.

VERB AGREEMENT: Presence versus absence of some kind of verb agreement. *Data source:* expanded dataset from Nichols & Bickel (2005) and Bickel & Nichols (2005b) in AUTOTYP.

POSSESSOR AGREEMENT: Presence versus absence of some kind of possessor agreement in noun phrases with open-class heads (e.g., not limited to inalienables). *Data source:* an expanded dataset from Nichols & Bickel (2005) in AUTOTYP.

GRAMMATICAL CASE: Presence versus absence of case or adpositions that differentiate between agentive and patientive lexical noun phrases of transitive clauses under at least some conditions (e.g., when definite). *Data source:* Comrie (2005), Witzlack-Makarevich et al. (2011) and Witzlack-Makarevich (2011).

SPLIT ERGATIVITY OF AGREEMENT MARKERS: Proportion of ergative alignments among all agreement markers and conditions of splits that a language may have, such as aspect or person (binned into equally spaced intervals: low, medium, high). *Data source*: Witzlack-Makarevich et al. (2011) and Witzlack-Makarevich (2011)

SPLIT ERGATIVITY OF CASE: Proportion of ergative case alignments among all conditions of splits that a language may have, such as aspect or person (binned into equally spaced intervals: low, medium, high). *Data source*: Witzlack-Makarevich et al. (2011) and Witzlack-Makarevich (2011)

POLYEXPONENCE: Presence versus absence of any kind of polyexponential markers in either argument-related (chiefly case and number) or verbal (chiefly tense and negation) morphology, that is of markers that realize more than one category (e.g., case together with number as in Russian dative singular *stol-u* 'to the table' versus dative plural *stol-am* 'to the tables', or tense and aspect together with verb agreement as in Spanish perfective past *am-ó* 's/he loved' and *am-é* 'I loved' versus imperfective *am-aba* 's/he/I used to buy / bought'). *Data source:* expanded dataset from Bickel & Nichols (2005a) in AUTOTYP.

INFLECTIONAL COMPACTNESS: Ratio between the number of categories expressed by regular (major, default class) verb inflection and the number of formatives used for this: cumulative if the ratio is larger than 1, distributive if the ratio is smaller than 1; separative (agglutinative) if the ratio is 1. *Data source:* expanded version of Bickel & Nichols (2005b).

SYNCRETISM: Presence versus absence of any kind of syncretism in case or agreement systems (where absence could also mean total absence of the relevant system). *Data source:* Baerman & Brown (2005a,b).

FLEXIVITY: Presence versus absence of any kind of lexically conditioned allomorphy, chiefly in the form of declension, possession, or conjugation classes but also considering allomorphy of negation and number markers. *Data source:* AUTOTYP, with definitions from Bickel & Nichols (2007).

VERBAL SYNTHESIS: Degree of inflectional synthesis of the verb (number of categories expressed), binned into equally spaced rank-order intervals: low, medium, high. *Data source:* expanded dataset from Bickel & Nichols (2005b).

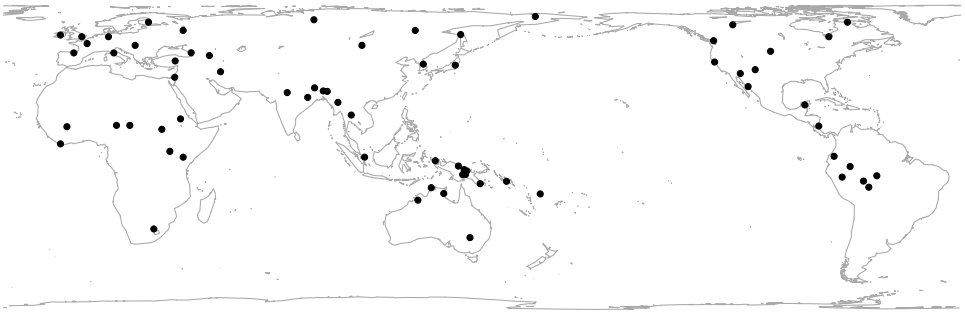
NOMINAL SYNTHESIS: Degree of inflectional synthesis of the noun, estimated from information about the presence of case, number, possessor agreement and definiteness affixes, ranging from 0 to 4, treated as categorically distinct levels. *Data source:* AUTOTYP and, for definiteness affixes, Dryer (2005a).

Some of these twelve variables, such as nominal synthesis, are only exploratory and need more detailed coding, and they are not all completely independent of each other. However, for present purposes, all we need is rough estimates of similarities that are relatively well grounded in empirical surveys. The variables we choose are not intended for precise measurements of distributions, let alone for categorical classifications of languages.

The first step in the analysis is to estimate the distance (dissimilarity) between each pair of languages. This can be easily done by computing the Hamming distance: the proportion of variables in which two languages have the same values. Estimating distances in this way requires that we have information for all variables. Missing information distorts the picture: for example, two pairs of languages may seem equidistant and have a dissimilarity value of .4, but in one pair this dissimilarity estimate could be based on four differences among ten variables, while in the other pair the same similarity could be based on two differences among five variables. In each case, the proportion is .4. If we are equally interested in all variables listed above, dissimilarity estimates need to be consistently based on the same range of variables. For a dozen languages, we had additional data readily available from reference grammars or handbooks, but this still leaves many gaps.

In response to this, we reduced the dataset by removing all languages that had missing information in at least one variable. This leaves us with only sixty-eight languages, but they cover the world relatively well (see Map 2 and the coding in the Appendix).

We then computed pairwise dissimilarities in the dataset and applied the fuzzy clustering algorithm of Kaufman & Rousseeuw (1990) to the data. In order to satisfy Constraint 1, that is, in order to arrice at a sample size realistic for acquisition research, we required the algorithm to group languages



Map 2. Languages with full data on the twelve variables discussed in the text

into no more than five clusters.[5] Figure 1 displays for each cluster all languages that have a membership coefficient at least twice as high as would be an equiprobable membership in the five clusters (which is .2). The languages with the highest coefficients are those that are the most prototypical representatives of the set (with area definitions from Nichols & Bickel 2009).

One of the clusters (Cluster 4) is well represented in the CHILDES database because it includes several European languages, while the other clusters are represented only marginally and Cluster 1 is not represented at all. These clusters urgently need more longitudinal corpora. Selecting languages for this in each cluster is a matter of practical considerations. Relevant criteria include the extent to which the language is still learned by children, whether native speaker assistants can be hired and motivated for transcription and translation work, general projects costs and similar concerns.

These practical considerations notwithstanding, the results in Figure 1 allow setting up a sample of five pairs that differ within themselves only minimally. Clearly, there are many other ways in which these languages in each cluster still differ from each other, and not all possible contrasts are captured (as this would require $\sum_{i=1}^{n} \binom{k_i}{2} = 30$ such contrast pairs and therefore a sample of sixty languages, given the variables analyzed here; see the discussion in Section 5). But for the variables of interest, the clustering in Figure 1 reduces these differences to the minimum that keeps the sample size small.

We have illustrated the method with variables relating to morphological complexity and word order. Needless to say, the method is completely neutral as to the variables entered, and, given suitable databases, one could just as well sample according to differences in the social conditions under which

[5] We used the facilities provided by Maechler et al. (2005) and set the membership exponent to $r = 1.3$, because the standard $r = 2$ required more clustering structure in the data than they would actually afford; see the discussion in Kaufman & Rousseeuw (1990) and Maechler et al. (2005).

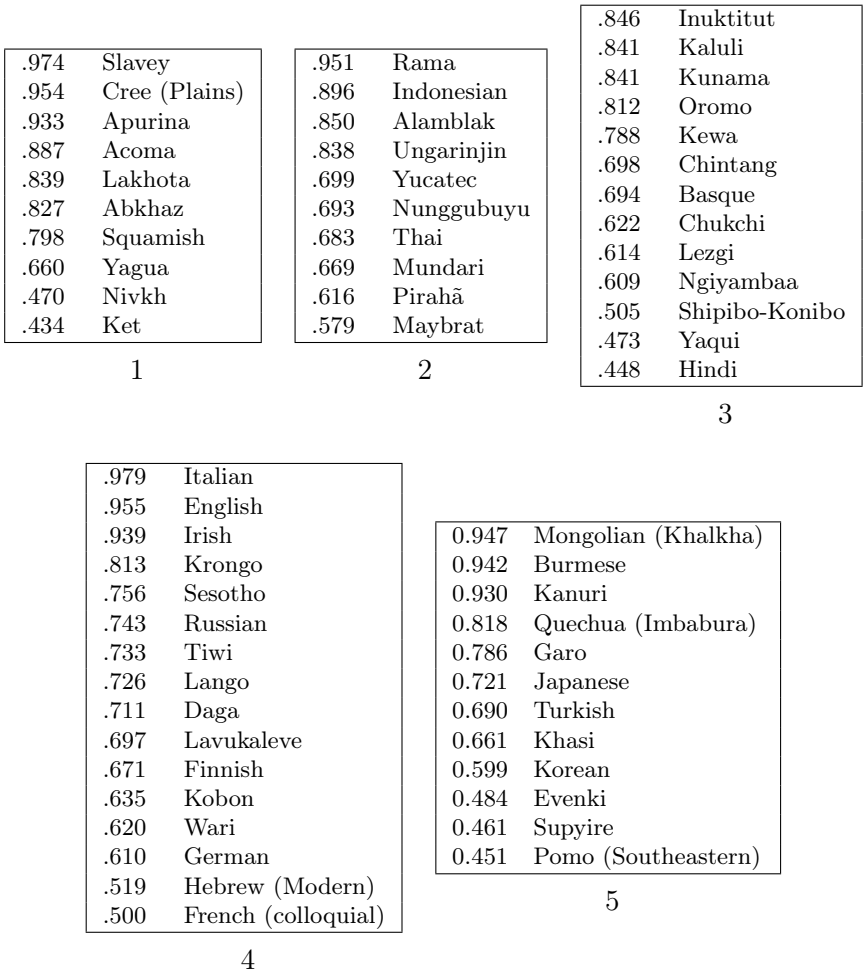


Figure 1. Fuzzy clusters based on the variables as defined in the text and tabulated in the Appendix, with membership coefficients in decreasing order

children learn languages (group network sizes, child-peer relations, role of parents, language teaching strategies, etc.) or - ideally - combine social and linguistic variables.

7. Discussion

The method proposed here yields language samples that satisfy the constraints formulated above: the best sample that one can set up, given the constraints. But we emphasize that the number of clusters is imposed onto the data - for the sole purpose of sampling. The clusters and their structure have no theoretical status of their own. Specifically, they cannot be construed as holistic types, that, is it would not make sense to postulate, for example, a Slavey or Rama prototype of languages that would represent an empirically motivated type of language. This becomes clear when examining the goodness-of-fit statistics for the clustering. A standard goodness-of-fit statistic for fuzzy clustering is known as the normalized Dunn coefficient, which ranges from 1 (crisp, well-motivated clustering) to 0 (no evidence for clustering, i.e., membership coefficients tend to be uniform across clusters) (see Kaufman & Rousseeuw 1990). The clustering shown in Figure 1 has a normalized Dunn coefficient of $D = .44$. This suggests that the clustering is to a considerable extent imposed onto the data and does not fall out naturally from the data.

The Dunn coefficient of a fuzzy cluster model depends on the kind and number of variables that are entered into the analysis and on the number of clusters that are requested. To find out whether the data support any natural crisp clustering, we computed the clusterings of all possible combinations of variables (with at least two and at most twelve variables; i.e., $n = \sum_{i=2}^{12} \binom{12}{i} = 4,083$ combinations) and a reasonable range of clusters (from two to eight).[6] The result of this is plotted in Figure 2. The gray scale indicates the highest normalized Dunn coefficient that was found per number of variables and number of clusters imposed onto the algorithm.

For combinations of many variables and few clusters, the normalized Dunn coefficient is relatively small ($D < .5$), suggesting strong fuzziness. This includes the design in our sampling study, which includes combinations that use all twelve variables and require five clusters. Higher coefficients can only be achieved with limited data (low values on the x-axis in Figure 2), but this is not very revealing because then the clusters simply mirror the categories of the variables, with little aggregation. Alternatively, coefficients can be increased by increasing the number of clusters (high values on the

[6] In some cases, the clustering algorithm did not converge and we treated these cases as providing even less evidence for a natural clustering than those with a small Dunn coefficient. The membership exponent was again set to $r = 1.3$, as noted in the case study.

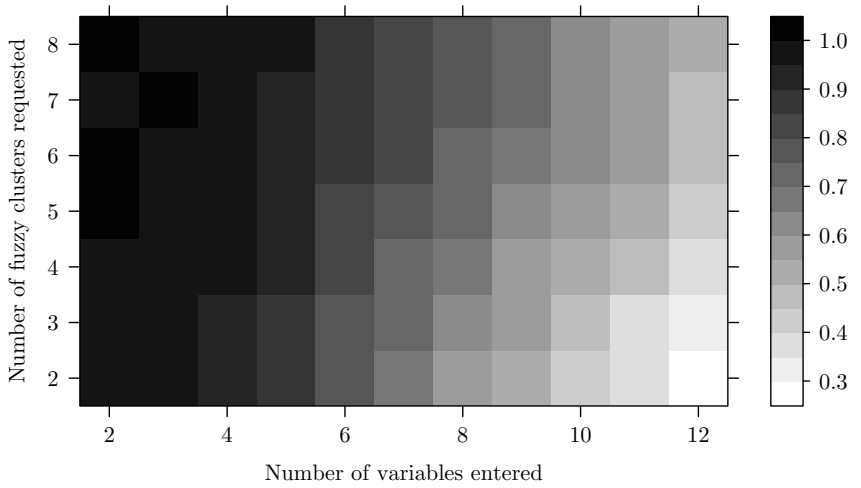


Figure 2. Degrees of highest crispness (normalized Dunn coefficient) achievable with all possible combinations of the twelve variables in the case study

y-axis) - but even then, coefficients become low again when many or all variables are entered.

We conclude from this that the kind of variables examined here do not define natural clusters of languages or prototypes. This suggests that despite partial dependencies between variables, they must have developed in their own ways to an extent that any covariation dissipates in larger samples. This is in line with one of the main findings of Nichols (1992): there is far less evidence for a small number of holistic language types than was assumed for a long time in typology.

8. Conclusions

Language acquisition research requires large longitudinal corpora. The development of these is extremely time consuming and expensive, and this raises the question of how one can still set up a sample of corpora that best captures the known structural diversity in the languages of the world. In response to this challenge, we propose to use fuzzy clustering algorithms on typological data. These algorithms allow one to derive samples that meet the critical constraints: samples can be relatively small but still capture diversity well and allow pairwise studies targeted at specific contrasts.

The clustered sampling method is neutral as to the kinds of diversity one wishes to consider. Apart from structural linguistic variables, social variables of language acquisition are of obvious importance. What makes work with these variables difficult, however, is the lack of sufficiently large

modern anthropological databases that track variation systematically and empirically responsibly.

Apart from the purposes we discussed here, the clustered sampling method is of general interest whenever one needs to identify those languages of a sample that differ from each other most. As such, the method is also ideally suitable for pilot studies when surveying some phenomenon across languages for the first time or when developing entirely new typological variables.

Appendix: Data used in the case study

ISO 639.3	Language	Stock	Verb pos.	Verb agr.	Poss. agr.	Case A vs.P	Agr. split erg.	Case split erg.	Polyexp.	Compactness	Syncr.	Flex.	V syn	N syn
abk	Abkhaz	West Caucasian	V=3	some	none	none	low	low	some	distributive	none	none	high	2
kjq	Acoma	Keresan	free	some	none	none	low	low	some	cumulative	none	some	medium	2
amp	Alamblak	Sepik	V=3	some	some	none	low	low	none	separative	none	none	high	1
apu	Apurina	Arawakan	V=2	some	none	none	low	low	some	distributive	none	some	medium	2
eus	Basque	Basque	V=3	some	some	some	high	high	none	distributive	some	none	low	3
mya	Burmese	Sino-Tibetan	V=3	none	some	some	low	low	some	separative	none	none	low	2
ctn	Chintang	Sino-Tibetan	V=3	some	some	some	low	medium	some	distributive	some	none	high	3
ckt	Chukchi	Chukchi-Kamchatkan	free	some	some	some	medium	high	some	distributive	some	some	medium	2
crk	Cree (Plains)	Algic	free	some	none	none	low	low	some	distributive	none	some	medium	2
dgz	Daga	Dagan	V=3	some	none	none	low	low	some	cumulative	some	some	medium	1
eng	English	Indo-European	V=2	some	some	none	low	low	some	cumulative	some	some	low	1
evn	Evenki	Tungusic	V=3	some	none	some	low	low	some	cumulative	none	none	medium	3
fin	Finnish	Uralic	V=2	some	none	some	low	low	some	distributive	some	some	low	3
fra	French (colloquial)	Indo-European	free	some	some	none	low	low	some	distributive	some	none	low	1
grt	Garo	Sino-Tibetan	V=3	none	some	some	low	low	none	separative	none	none	low	2
deu	German	Indo-European	V=2	some	some	some	low	low	some	cumulative	some	some	low	2
grb	Grebo	Kru	V=2	none	some	none	low	low	some	distributive	none	some	medium	1
hau	Hausa	Chadic	V=2	some	some	none	low	low	some	distributive	none	some	medium	1
heb	Hebrew (Modern)	Semitic	V=2	some	some	some	low	low	some	distributive	some	some	low	3
hin	Hindi	Indo-European	V=3	some	none	some	medium	medium	some	cumulative	some	some	low	2
hun	Hungarian	Uralic	V=2	some	none	some	low	low	some	cumulative	none	none	medium	3
imn	Imonda	Border	free	some	none	some	low	low	none	separative	none	some	high	3
ind	Indonesian	Austronesian	V=2	none	some	none	low	low	none	separative	none	none	low	1
iku	Inuktitut	Eskimo-Aleut	V=3	some	some	some	low	high	some	distributive	some	none	medium	3
gle	Irish	Indo-European	V=1	some	none	none	low	low	some	cumulative	some	some	low	1
ita	Italian	Indo-European	V=2	some	none	none	low	low	some	cumulative	some	some	low	1
jpn	Japanese	Japanese	V=3	none	some	some	low	low	some	cumulative	none	none	low	1
bco	Kaluli	Bosavi	V=3	some	some	some	low	medium	some	cumulative	some	some	medium	2
kau	Kanuri	Saharan	V=3	some	some	some	low	low	some	separative	none	none	medium	2
ket	Ket	Yeniseian	V=3	some	none	none	medium	low	some	distributive	some	some	low	2
kew	Kewa	Engan-Kewa	V=3	some	some	some	low	high	some	cumulative	some	none	medium	1
kha	Khasi	Austroasiatic	V=2	some	some	some	low	low	some	distributive	none	none	low	2
kpw	Kobon	Madang	V=3	some	some	none	low	low	some	cumulative	some	some	medium	1
kor	Korean	Korean	V=3	none	some	some	low	low	some	distributive	none	some	medium	2

Appendix (*continued*)

ISO 639.3	Language	Stock	Verb pos.	Verb agr.	Poss. agr.	Case A vs.P	Agr. split erg.	Case split erg.	Polyexp.	Compactness	Syncr.	Flex.	V syn	N syn
kgo	Krongo	Kadugli-Krongo	V=1	some	some	none	low	low	some	distributive	some	some	low	1
kun	Kunama	Kunama	V=3	some	some	some	low	low	none	cumulative	some	some	medium	3
lkt	Lakhota	Siouan	V=3	some	none	none	medium	low	some	distributive	none	some	high	2
laj	Lango	Nilotic	V=2	some	none	none	low	low	some	cumulative	some	some	medium	3
lvk	Lavukaleve	Central Solomons	V=3	some	none	none	low	low	some	cumulative	some	some	low	2
lez	Lezgi	Nakh-Daghestanian	V=3	none	some	some	low	high	none	distributive	some	some	low	2

Appendix (*continued*)

ISO 639.3	Language	Stock	Verb pos.	Verb agr.	Poss. agr.	Case A vs.P	Agr. split erg.	Case split erg.	Polyexp.	Compactness	Syncr.	Flex.	V syn	N syn
mrc	Maricopa	Yuman	V=3	some	none	some	low	low	none	distributive	none	some	medium	3
ayz	Maybrat	West Papuan	V=2	some	some	none	low	low	some	separative	none	some	low	0
khk	Mongolian (Khalkha)	Mongolic	V=3	some	some	some	low	low	some	separative	none	none	low	2
muw	Mundari	Austroasiatic	V=3	some	some	none	low	low	none	distributive	none	some	medium	1
wyb	Ngiyambaa	Pama-Nyungan	V=3	some	some	some	medium	medium	none	cumulative	some	some	low	1
niv	Nivkh	Nivkh	V=3	some	none	none	low	low	none	distributive	some	none	medium	2
nuy	Nunggubuyu	Gunwingguan	free	some	some	none	low	low	none	separative	some	some	low	1
hae	Oromo	Cushitic	V=3	some	some	some	low	low	none	cumulative	some	some	medium	2
pes	Persian	Indo-European	V=3	some	none	some	low	low	some	separative	none	some	low	1
myp	Pirahã	Muran	V=3	none	some	none	low	low	none	separative	none	none	medium	0
pom	Pomo (Southeastern)	Pomoan	V=3	none	some	some	low	low	none	distributive	none	some	medium	2
que	Quechua (Imbabura)	Quechuan	V=3	some	some	some	low	low	some	distributive	none	none	medium	2
rma	Rama	Chibchan	V=3	some	some	none	low	low	none	separative	none	some	low	1
rus	Russian	Indo-European	V=2	some	none	some	low	low	some	cumulative	some	some	low	2
smo	Samoan	Austronesian	V=1	none	none	some	low	medium	none	separative	none	none	low	1
sot	Sesotho	Benue-Congo	V=2	some	none	none	low	low	some	distributive	some	none	low	1
shp	Shipibo-Konibo	Pano-Tacanan	V=3	some	some	some	low	high	none	cumulative	none	none	medium	2
scs	Slave	Na-Dene	V=3	some	none	none	low	low	some	distributive	none	some	medium	2
squ	Squamish	Salishan	V=1	some	none	none	low	low	some	cumulative	none	none	medium	2
spp	Supyire	Senufo	V=3	none	none	none	low	low	some	separative	none	none	low	3
tha	Thai	Tai-Kadai	V=2	none	some	none	low	low	none	distributive	none	none	low	0
tiw	Tiwi	Tiwi	V=2	some	some	none	low	low	some	distributive	some	some	medium	1
tur	Turkish	Turkic	V=3	some	none	some	low	low	some	separative	none	none	medium	3
ung	Ungarinjin	Worrorran	V=2	some	some	none	medium	low	none	separative	none	some	high	1
pav	Wari	Chapakuran	V=1	some	none	none	low	low	some	cumulative	none	some	low	1
yad	Yagua	Yaguan	V=1	some	none	none	medium	low	some	cumulative	none	none	high	2
yaq	Yaqui	Uto-Aztecan	V=3	some	none	some	low	low	some	distributive	some	none	medium	3
yua	Yucatec	Mayan	free	some	some	none	medium	low	none	separative	none	some	low	2

References

van der Auwera, Johan. 1998. Conclusion. In Johan van der Auwera (ed.), *Adverbial constructions in the languages of Europe*, 813–836. Berlin: Mouton de Gruyter.

Baerman, Matthew & Dunstan Brown. 2005a. Case syncretism. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 118–121. Oxford: Oxford University Press.

Baerman, Matthew & Dunstan Brown. 2005b. Syncretism in verbal person/number marking. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 122–125. Oxford: Oxford University Press.

Bakker, Dik. 2011. Language sampling. In Jae Jung Song (ed.), *The Oxford Handbook of Language Typology*, 90–127. Oxford: Oxford University Press.

Behrens, Heike. 2002. Learning multiple regularities: Evidence from overgeneralization errors in the German plural. In Anna H.-J. Do, Laura Domínguez & Aimee Johansen (eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development*, 61–71. Somerville, MA: Cascadilla Press.

Bell, Alan. 1978. Language samples. In Joseph H. Greenberg, Charles Ferguson & Edith Moravcsik (eds.), *Universals of human language I: Method and theory*, Stanford: Stanford University Press.

Berko Gleason, Jean. 1958. The child's learning of English morphology. *Word* 14. 150–177.

Berman, Ruth A. & Dan I. Slobin. 1994. *Relating events in narrative: A crosslinguistic developmental study.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bickel, Balthasar. 2008. A refined sampling procedure for genealogical control. *Language Typology and Universals* 61. 221–233.

Bickel, Balthasar & Johanna Nichols. 2005a. Exponence of selected inflectional formatives. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 90–93. Oxford: Oxford University Press.

Bickel, Balthasar & Johanna Nichols. 2005b. Inflectional synthesis of the verb. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 94–97. Oxford: Oxford University Press.

Bickel, Balthasar & Johanna Nichols. 2007. Inflectional morphology. In Timothy Shopen (ed.), *Language typology and syntactic description*, 169–240. Cambridge: Cambridge University Press (revised second edition).

Borer, Hagit & Kenneth Wexler. 1987. The maturation of syntax. In Thomas Roeper & Edwin Williams (eds.), *Parameter-setting and language acquisition*, 23–172. Dordrecht: Reidel.

Bowerman, Melissa & Stephen C. Levinson. 2001. *Language acquisition and conceptual development.* Cambridge: Cambridge University Press.

Choi, Soonja. 2000. Caregiver input in English and Korean: Use of nouns and verbs in book-reading and toy-play contexts. *Journal of Child Language* 27. 69–96.

Choi, Soonja & Melissa Bowerman. 1991. Learning to express motion events in English and Korean - the influence of language-specific lexicalization patterns. *Cognition* 41. 83–121.

Chomsky, Noam. 1968. *Language and mind.* New York: Harcourt Brace Jovanovich.

Chomsky, Noam. 1980. *Rules and representations.* Oxford: Blackwell.

Chomsky, Noam. 1981. Principles and parameters in syntactic theory. In Norbert Hornstein & David Lightfoot (eds.), *Explanation in linguistics: The logical problem of language acquisition*, 32–75. London: Longman.

Comrie, Bernard. 2005. Alignment of case marking. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 398–405. Oxford: Oxford University Press.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257–292.

Dryer, Matthew S. 2005a. Definite and indefinite articles. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 154–161. Oxford: Oxford University Press.

Dryer, Matthew S. 2005b. Order of subject, object, and verb. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 330 – 334. Oxford: Oxford University Press.

Evans, Nicholas & Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32. 429–448.

Everett, Daniel L. 2005. Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology* 46(4). 621–646.

Everett, Daniel L. 2009. Pirahã culture and grammar: a response to some criticism. *Language* 85. 405–442.

Fitch, W. Tecumseh, Marc D. Hauser & Noam Chomsky. 2005. The evolution of the language faculty: Clarifications and implications. *Cognition* 97. 179–210.

Gentner, Dedre. 1982. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In Stanley A. Kuczaj (ed.), *Language development*, vol. 2, 38–62. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gentner, Dedre & Lera Boroditsky. 2001. Individuation, relativity, and early word learning. In Melissa Bowerman & Stephen C. Levinson (eds.), *Language acquisition and conceptual development*, 215–256. Cambridge: Cambridge University Press.

Haspelmath, Martin. 1998. How young is Standard Average European? *Language Sciences* 20. 271–287.

Haspelmath, Martin. 2001. The European linguistic area: standard average European. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook*, 1492–1510. Berlin: Mouton de Gruyter.

Hyams, Nina. 1986. *Language acquisition and the theory of parameters.* Norwell, MA: Reidel.

Kaufman, Leonard & Peter J. Rousseeuw. 1990. *Finding groups in data: An introduction to cluster analysis.* New York: Wiley.

Köpcke, Klaus-Michael. 1998. The acquisition of plural marking in English and German revisited: schemata versus rules. *Journal of Child Language* 25. 293–319.

Ladd, D. Robert, Bert Remijsen & Caguor Adong Manyang. 2009. On the distinction between regular and irregular inflectional morphology: Evidence from Dinka. *Language* 85. 659–670.

Levinson, Stephen C. 2003. *Space in language and cognition.* Cambridge: Cambridge University Press.

Lieven, Elena V. M. & Sabine Stoll. 2010. Language. In Marc H. Bornstein (ed.), *Handbook of cultural developmental science*, 143–160. Psychology Press.

Maechler, Martin, Peter J. Rousseeuw, Anja Struyf & Mia Hubert. 2005. `cluster`: cluster analysis basics and extensions. `R` package, `http://www.R-project.org/`.

Nichols, Johanna. 1992. *Linguistic diversity in space and time.* Chicago: The University of Chicago Press.

Nichols, Johanna. 1996. The Comparative Method as heuristic. In Mark Durie & Malcolm Ross (eds.), *The Comparative Method reviewed*, 39–71. Oxford: Oxford University Press.

Nichols, Johanna. 2008. Linguistic complexity: A comprehensive definition and survey. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 110–125. Oxford: Oxford University Press.

Nichols, Johanna & Balthasar Bickel. 2005. Locus of marking (in the clause; in possessive noun phrases; and whole-language typology). In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 98–109. Oxford: Oxford University Press.

Nichols, Johanna & Balthasar Bickel. 2009. The AUTOTYP genealogy and geography database: 2009 release. Electronic database, `http://www.uzh.ch/spw/autotyp`.

Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13. 293–315.

Pinker, Steven. 1994. *The language instinct.* New York: Morrow.

Pinker, Steven & Michael Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Science* 6. 456–463.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2. 263–314.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & Peter Kahrel. 1993. A method of language sampling. *Studies in Language* 17. 169–203.

Slobin, Dan I. 1973. Cognitive prerequisites for the development of grammar. In Charles A. Ferguson & Dan I. Slobin (eds.), *Studies in child language development*, 175–208. New York, NY: Holt, Rinehart, & Winston.

Slobin, Dan I. 1985. Introduction: why study acquisition crosslinguistically? In D. I. Slobin (ed.), *The crosslinguistic study of language acquisition Volume 1: The data*, 3–24. Hillsdale, NJ: Erlbaum.

Stoll, Sabine. 1998. The role of Aktionsart in the acquisition of Russian aspect. *First Language* 18. 351–377.

Stoll, Sabine. 2005. Beginning and end in the acquisition of the perfective aspect in Russian. *Jounal of Child Language* 32. 805–825.

Stoll, Sabine. 2009. Crosslinguistic approaches to language acquisition. In Edith L. Bavin (ed.), *The Cambridge handbook of child language*, 89–104. Cambridge: Cambridge University Press.

Stoll, Sabine, Balthasar Bickel, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Judith Pettigrew, Ichchha P. Rai, Manoj Rai & Novel Kishore Rai. 2012. Nouns and verbs in Chintang: children's usage and surrounding adult speech. *Journal of Child Language* 39. 284–321.

Tardif, Twila, Marilyn Shatz & Letitia R. Naigles. 1997. Caregiver speech and children's use of nouns versus verbs: a comparison of English, Italian, and Mandarin.

Journal of Child Language 24. 535–565.

Tardif, Twila. 1996. Nouns are not always learned before verbs: evidence from Mandarin speakers' early vocabularies. *Developmental Psychology* 32. 492–504.

Van Valin, Robert D., Jr. 2009. Some remarks on Universal Grammar. In Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura & Seyda Özçalişkan (eds.), *Crosslinguistic approaches to the psychology of language: research in the traditions of Dan Slobin*, 311–320. New York: Psychology Press.

Witzlack-Makarevich, Alena. 2011. *Typological variation in grammatical relations.* Leipzig: University of Leipzig dissertation.

Witzlack-Makarevich, Alena, Lennart Bierkandt, Taras Zakharko & Balthasar Bickel. 2011. AUTOTYP database on grammatical relations. Electronic database, University of Leipzig [`www.uzh.ch/spw/autotyp`].