Data reduction typology and the bimodal distribution bias

BERNHARD WÄLCHLI

Abstract

Confronting low data reduction typologies, as established by using data from parallel texts, with the high data reduction typologies of WALS *reveals a systematic bias of* WALS *typologies toward highly bimodal distribution. Properties with a distribution supporting a discrete feature analysis in many languages are likelier to be represented in* WALS *and to be represented accurately. This bias has important consequences when* WALS *typologies are interpreted theoretically or further processed statistically.*

Keywords: *data reduction, doculect, imperative, linguistic atlas, methodology, number, parallel texts, word order*

1. Introduction

Typological investigations traditionally show a preference for strong data reduction. Data reduction is any summarizing or grouping of data where information might be lost. If your water gage measures on three occasions A 1.03 m, B 1.97 m, and C 5.19 m, the data is reduced to various extents if you take down instead A 1 m, B 2 m, C 5 m; or A low, B middle, C high; or A low, B low, C high. *WALS* practices strong data reduction systematically. An extreme case is Maddieson's (2005) data on consonant inventory size, which is reduced to a scale of small (6–14), moderately small (15–18), average (19–25), moderately large (26–33), and large (34 or more), even though Maddieson's original source, Maddieson 1984, contained the exact size of consonant phonemes per language.

The common typological practice of data reduction is connected with the dominant source of information: reference grammars. A first step is taken by the grammar writer who applies data reduction in using general descriptive terms. Constituent order, for instance, can be measured only in individual ex-

Linguistic Typology 13 (2009), 77–94
DOI 10.1515/LITY.2009.004
1430–0532/2009/013-077

amples. Every statement about dominant or preferred order in the constructions of a language implies strong data reduction. In the next step, the typologist further reduces the level of granularity to a level covered by a large number of reference grammars. For instance, Turkish is simply classified as a language with dominant object-verb order by Dryer (2005d) – as is common practice in word order typology – despite statements such as the following in the grammars concerned (Lewis 1967: 243):

> In the colloquial, an imperative often begins a sentence, because someone with urgent instructions to give will naturally put the operative word first: *çık oradan!* 'get out of there!' *yakma lâmbayı!* 'don't light the lamp!'

Data reduction is unavoidable if large amounts of information are surveyed. However, it is important for methodological reasons to be aware of its twin: loss of information. Intriguing questions are: How much relevant information is lost? Does data reduction entail certain kinds of systematic errors? Are there any biases due to extreme data reduction such as practiced in *WALS*? The aim of this article is to show that the kinds of distributions of typological features observable only at a lower level of data reduction have consequences for a typological database containing features with strong data reduction, notably the following:

(i) Features with bimodal distributions (with extreme poles) are intuitively preferred and hence overrepresented in the database (bimodal distribution bias) because a discrete classification is adequate for a large number of languages. Hence, bimodally distributed features tend to be accurately represented by discrete features as a whole even though discrete features are not accurate for all languages.

(ii) Features without polar distributions are underrepresented in typological databases such as *WALS* and those which are covered are less accurately represented by discrete features.

It might be argued that the discussion here confounds competence (or *langue*) and performance (or *parole*) and that typological databases should be about competence and not performance, which justifies a coarse level of data representation. However, language data is always performance, and competence can only be approached by modeling which components of performance are relevant. In empirical terms, languages as systems cannot be described directly, and as Halliday (1991: 42) puts it: "frequency in text is the instantiation of probability in the system". Because it is not *a priori* clear how representative an available documentation is for a particular language, it is preferable to adopt the notion of doculect. The term DOCULECT has been coined by Michael Cysouw, Jeffrey Good, and Martin Haspelmath in 2006 at the Max Planck Institute for Evolutionary Anthropology to denote a variety of a language that has been described or otherwise documented. It is first mentioned in the literature

by Bowern (2008: 8). A doculect can be any documented language variety, be it as raw data (e.g., sound file), primary data (e.g., a transcribed text), or secondary data (description, e.g., a reference grammar) of whatever size. The doculect is related to language as sample to population in statistics. A doculect can thus be more or less representative of a language.

In Section 2 two *WALS* typologies with different underlying frequency distributions are compared to datasets from parallel texts which make it possible to assess the different nature of the underlying crosslinguistic frequency distributions and to study the interaction of distribution and accuracy in a typology with strong data reduction. In Section 3 it is argued that discrete categories obtained by means of strong data reduction do not support an underlying parametric structure of grammar.

2. Partial matching of two *WALS* typologies with exemplar-based datasets from parallel texts

2.1. *Reciprocal and partial evaluation and distorting "universals of translation"*

Empirical work should be evaluated, and the best way to do this is by means of other related empirical work based on different sources of material which is in need of evaluation itself. Every type of data source in typology has different kinds of advantages and disadvantages, which is why empirical evaluation is always reciprocal. Furthermore, since two different typological datasets will hardly ever cover exactly the same range of languages, we have to deal here with partial evaluation. Partial evaluation is the more useful the more the intersection of languages considered is representative of the full sample.

In this section I consider a typology with bimodal distribution, order of object and verb (Dryer 2005d), and a typology with a distribution lacking salient frequency poles, occurrence of nominal plurality (Haspelmath 2005). It is shown that the former is more accurate, not because Dryer worked more carefully than Haspelmath, but because the underlying distributions are different. In order to show this, I adduce other datasets that allow me (i) to determine the nature of the underlying distributions at least approximately and (ii) to match those datasets with the *WALS* typologies at least partly so that it can be assessed to a certain extent what happens in the discrete *WALS* classifications in which ranges of the continuous distributions.

In order to get datasets with a low degree of data reduction we must consider texts, ideally texts which are comparable at the level of individual clauses and available in many languages. Such texts are massively parallel texts (Cysouw & Wälchli (eds.) 2007), texts translated into many languages. One of few texts suitable for this purpose is the Gospel according to Mark, which is being used here. This comes at a cost: translations are not fully representative for

languages. (However, many reference grammars also rely on translated texts to a larger or smaller extent, but this has apparently never been a major source of concern for typologists.)

Translations do not behave exactly like original texts, but deviate from them in ways not fully unpredictable. Scholars of translation studies speak of "universals" of translation (Mauranen & Kujamäki 2004), which means that there are systematic tendencies in the way that at least some translated texts deviate from original texts. Some of these general tendencies in translation do not do any harm for our purposes. For instance, many translations have a tendency toward explication, which would mean that they have more clauses. This tendency is eliminated by considering only a clearly defined set of clauses or nouns of the source text with their direct translation equivalences. The two most relevant remaining systematic tendencies of translation are the following:

(i) "[G]rowing grammatical conventionality and a tendency to overrepresent typical features of the target language" (Mauranen & Kujamäki 2004: 1). This means, for instance, that word order is expected to be more rigid (less free) in translations than in original texts.

(ii) Interference from the source language resulting in a greater structural similarity with the source language than original texts would have. Here it has to be taken into account that there are different source languages in Bible translations. However, as far as word order is concerned, the dominant source languages – Classical Greek, Latin, Spanish, English, Russian, and Indonesian – all happen to have dominant VO order which is why the expected effect of (ii), if it is at work, will be a general overrepresentation of VO.

These tendencies will have to be taken into account in the concrete datasets considered below.

The typologies compared should ideally cover exactly the same domains and differ only in the source of data (reference material vs. texts) and in the data mode (discrete vs. frequency count). If this is the case, they are expected to show an ideal match. In case of an ideal match the discrete classes should exactly correspond to clearly definable segments of the frequency scale. However, such an ideal match cannot be expected since the comparison is problematic in several respects which will have to be discussed below: (i) comparable data are available only for a subset of languages, and in few cases closely related languages expected to behave similarly are included; (ii) the compared typologies do not cover exactly the same domain; (iii) the values in the parallel text data are expected to be distorted by translation in certain systematic ways; and (iv) the counts cover only small corpora from a single text. However, the main issue is not to find ideal matches but to explore whether or not there is a very close match, where some few remaining exceptions can be explained away when the data is considered more closely.

2.2. *The two WALS datasets considered*

Two discrete *WALS* datasets, Order of object and verb (Dryer 2005d, *WALS* Map 83), and Occurrence of nominal plurality (Haspelmath 2005, *WALS* Map 34), are compared to frequency counts in parallel texts.

Dryer's typology has three values, "OV", "VO", and "No dominant order". Dominance is explicitly connected to frequency in the first constituent order chapter (Dryer 2005c: 371):

> In some languages with flexible order, there is one order which is most common and which can be described as the dominant order. In other flexible order languages, the flexibility is greater and there is no one order that is the dominant order in terms of frequency of usage or pragmatic neutrality.

Thus, it does not seem to be unfair to compare it to actual frequency counts. Haspelmath's typology has six values which arise as the combination of two related typologies: occurrence of nominal plurality (i) in human nouns and (ii) in inanimate nouns. For both of them there are the three values "no", "optional", and "obligatory" related by an implicational universal, which is why three of nine possible combinations are not attested. Even though "extent to which plural markers on full nouns are used" is framed in modal terms rather than in terms of frequency it is expected to fully correlate with frequency: "no" occurrence is expected to be less frequent than "optional" occurrence, and "optional" occurrence less frequent than "obligatory" occurrence.

2.3. *The Mark order dataset*

The Mark order dataset derives from the order of local phrases and main verbs in 190 motion event clauses aligned across 100 translations of Mark from languages from all continents (hence 19,000 datapoints).[1] The database is a by-product of a study on motion events (Wälchli in preparation). The sample

1. Sample: Africa (18 languages): Bari, Ewe, Gbeya Bossangoa, Hausa, Ijo (Kolokuma), Ju|'hoan, Kabba-Laka, Kabyle, Khoekhoe, Koyra Chiini, Kunama, Maltese, Moru, Murle, Nubian (Kunuz), Somali, Swahili, Wolof; Creoles (2 languages): Papiamentu, Tok Pisin; Eurasia (17 languages): Adyghe (Temirgoy), Ainu, Avar, Basque, Breton, Georgian, Greek (Classical), Hindi, Kannada, Khalkha, Korean, Lak, Lezgian, Liv, Mansi, Mari (Meadow), Tuvan; South East Asia and Oceania (14 languages): Garo, Hmong Njua, Jabêm, Khasi, Lahu, Malagasy, Mandarin, Maori, Mizo, Nicobarese (Car), Santali, Thai, Timorese, Vietnamese; New Guinea and Australia (20 languages): Ambulas, Arapesh, Burarra, Daga, Enga, Kala Lagaw Ya, Kâte, Kiwai, Kuku-Yalanji, Kuot, Motuna, Nunggubuyu, Pitjantjatjara, Sougb, Toaripi, Tobelo, Waris, Warlpiri, Wik Munkan, Worora; North and Mesoamerica (16 languages): Cakchiquel, Choctaw, Comanche, Cree (Plains), Dakota, Hopi, Mixe (Coatlán), Mixtec (San Miguel el Grande), Navajo, Otomí (Mezquital), Purépecha, Tarahumara (Western), Tlapanec, Totonac (Sierra), Zapotec (Isthmus), Zoque (Copainalá); South America (13 languages): Amuesha, Aymara, Bribri, Chiquito, Guaraní, Kuna, Mapudungun, Miskito, Ngäbere, Paumarí, Piro, Quechua (Imbabura), Shipibo-Konibo. 53 families and 91 genera

was chosen so as to cover as much linguistic diversity as possible, genealogically and areally, within the given restriction of translations of Mark available to me. No well-established language family is represented with more than five languages, "phyla" with more than five are Australian (eight) and Nilo-Saharan (seven).The domain is slightly different from Dryer 2005d since "local phrases" are considered instead of objects. The term "local phrase" (L) denotes here any nominal, adverbial, or pronominal expression of the ground in motion events (semantic roles of goal, source, and companion), be it marked by an adposition and/or case or be it unmarked. As is common in word order typology, this is a functional domain rather than a formal category. It is well-known that oblique phrases tend to be on the same side of the verb as the direct object in most languages, which is why Hawkins (2004: 124) uses Dryer's object verb data as direct evidence for the explanation of V-PP order. Dryer & Gensler (2005) mention thirty-seven languages with OVX and three with XVO order ("X" is their label for oblique), the latter all Sinitic (Mandarin, Cantonese, Hakka). Since Mandarin has dominant VL order in motion events at least in some texts including Mark, XVO can be completely disregarded. However, OVX languages will cause mismatches in the evaluation due to differences in domain rather than in data source.

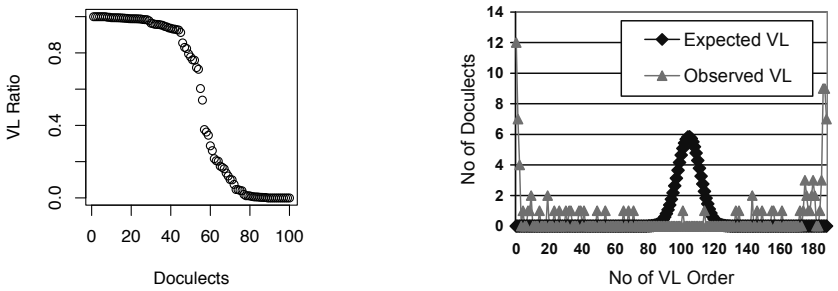
Figure 1 shows that the distribution of the verb-local phrase (VL) ratio in the 100 language sample is strongly bimodal: most doculects have either a very high or a very low VL ratio, with very few doculects around 50 %. However, this does not mean that all doculects have rigid order. In many doculects there are some exceptions, but if compared to a random distribution with the same overall probability of VL order (Figure 1b, assuming that a binomial distribution would be random, see also Cysouw 2002), only two doculects, Kala Lagaw Ya and Wik Mungkan, both Pama-Nyungan, fall in the frequency range where all languages would have to be expected in a binomial distribution. This means that most non-rigid order languages have an order preference. Thus, non-rigid order does not mean free order.

Bimodal distributions have the advantage that at least for the languages at the extreme poles little information is lost if the continuous ratio is reduced to the mode (most frequent value) or cut in two segments at any other non-arbitrary or arbitrary cut-off point. However, the less reduced data can still be relevant for a better understanding of the crosslinguistic behavior of constituent order, since there are many languages that are not strictly rigid.

Counting frequencies is also strong data reduction, even though it is less

---

according to the *WALS* classification. Genera represented with more than one language are all genealogically rather diverse: Pama-Nyungan (5 languages), Arawakan (2 languages), Finnic (2 languages), Creoles and Pidgins (2 languages), Mixe-Zoque (2 languages), Oceanic (2 languages).



(a) VL ratio in the 100 doculects in ascending order

(b) Observed and expected random (binomial) distribution

Figure 1. *The distribution of VL order ratio in 190 passages encoding motion events in translations of Mark in 100 languages from all continents*

strong than forming two discrete classes. I would like to give just one more example of what kind of information can be found if the original distribution on the exemplar-level is retained in the database. We have seen above that imperatives and non-imperatives can behave differently concerning constituent order. We can now go through all 190 contexts and split them into an imperative (n=20) and a non-imperative (n=170) domain and compare the VL ratio for the two domains across all doculects. Figure 2 plots the VL ratio in the imperative domain on the y-axis against the non-imperative domain on the x-axis. The languages of the 100 language sample are plotted in black, some additional languages are given in gray. On the basis of the values of the 100 language sample, a linear tail-restricted cubic spline function with five knots (Harrell 2001, function rcs()) of non-imperative is fitted with imperative in a linear regression model (Harrell 2001, function ols()) with a 95 % confidence interval (dashed lines), which shows that there is a significant universal tendency toward VL in imperative contexts in flexible order languages with dominant LV order. (The dotted line indicating equal VL ratios is not within the confidence interval for lower VL ratio levels.) This tendency is strongest in some doculects of languages of Eurasia, such as Basque, Lak, Udmurt, Lezgian, and Ossetic, but there is no language with rigid VL order in the imperative domain and rigid LV order in the non-imperative domain. Put differently, the tendency toward imperative first rather than non-imperative first is nearly universal in discourse, but it is never fully grammaticalized in any language since there are no rigid VL imperative & LV non-imperative languages. Accordingly, it tends to escape grammatical description which is biased toward fully grammaticalized structures.

Typologists have a genuine interest in general tendencies whether universal

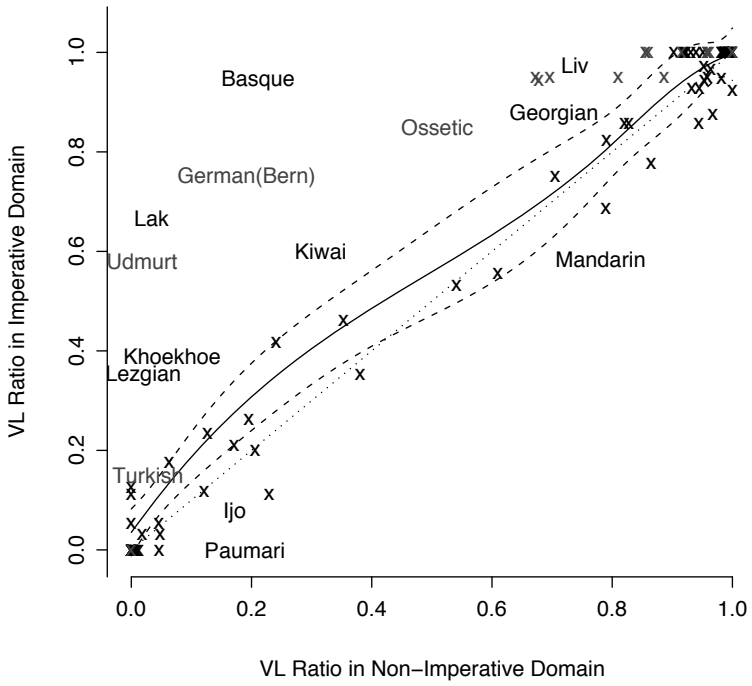


Figure 2. *The Imperative factor (only some doculects labeled, others indicated by "x")*

or statistic. Thus a finding such as the one supported by Figure 2, that imperatives tend to have VL order more often than non-imperatives, is relevant for constituent order typology. However, this finding could never be made in the strong data reduction mode of *WALS*. The most relevant issue here is really the degree of data reduction whatever the data source. The parallel text data suggest that the imperative-first tendency is particularly strong in Basque, and indeed some sources on Basque grammar note it ("Rule 3: A finite verb form [...] must not stand at the beginning of a sentence [...] Rule 3a: Rule 3 is waived for imperatives, in which the verb typically comes first", King & Olaizola Elordi 1996: 203), but many others do not. On the other hand, Turkish is almost rigid LV in the Mark translation (15 % LV ratio in the imperative domain), testifying to the "universal of translation" of growing grammatical conventionality. OV and LV orders are the norm of written Turkish and as such are implemented also in the translation of Mark as in many other original texts in written Turkish. This is important for our purposes in more general terms. If translations are often subject to "growing grammatical conventionality" this means for constituent order that word order in translations is expected to be

slightly more rigid (more strongly bipolar) than in colloquial language. As a consequence, the constituent order is expected to be less bipolar in original texts but will still be bipolar, since translation only reinforces a tendency, but it does not create it.

2.4. *The Mark number dataset*

The Mark number dataset derives from a database of all aligned instances of non-singular nominal number (including plural, dual, paucal, etc.) in Mark 8 with a literal equivalence in the Greek original text (to avoid the effect of higher degree of explication in translations). The sample used is a convenience sample of 84 languages but contains languages from all continents.[2]

The frequency distribution of non-singular nominal number is not bimodal in the Mark number data set (Figure 3). There are few languages lacking nominal number and there is no absolute maximum. Languages with a high frequency of nominal non-singular markers tend to have pluralia or dualia tantum (see Koptjevskaja-Tamm & Wälchli 2001: 629–637) with an idiosyncratic distribution. The set of languages lacking nominal number does not match completely with Haspelmath 2005 and Dryer 2005a, but in all studies the proportion of languages completely lacking nominal number is small (9.7 % in Haspelmath 2005 and 9.0 % in Dryer 2005a). While the data considered here does not allow us to determine the exact typological frequency distribution of nominal number, it is clear that the distribution is not strongly bimodal and lacks clear cut-off points. Contrasting the distribution with a random binomial distribution does not make much sense. However, it becomes obvious that the distribution of nominal number markers is not random if the frequency of particular semantic groups is considered. Figure 3 shows the frequency of non-singular in six semantic groups which form a characteristic semantic profile. Low total frequency goes together with a high proportion of nouns for humans marked for non-singular and high total frequency goes together with a higher proportion of body parts and explicitly quantified NPs marked for non-singular (e.g., *how many baskets*, *five loaves*), and of pluralia tantum.

---

2. The sample: Ainu, Akan, Amuesha, Avar, Aymara, Bambara, Basque, Bribri, Cakchiquel, Chamorro, Chiquito, Choctaw, Chuvash, Dakota, Drehu, Dungan, Efik, English, Estonian, Ewe, Finnish, French, Garo, Georgian, Greek (Classical), Guaraní, Haitian Creole, Hawaiian, Hmong Njua, Hungarian, Igbo, Indonesian, Jamaican Creole, Ju|'hoan, Kâte, Khalkha, Khasi, Khoekhoe, Kiwai, Korean, Koyra Chiini, Kuna, Kuot, Lahu, Latvian, Lezgian, Lithuanian, Malagasy, Maltese, Mandarin, Mapudungun, Mari (Meadow), Miskito, Mixe (Coatlán), Mordvin (Erzya), Nicobarese (Car), Nubian (Kunuz), Nunggubuyu, Ossetic, Otomí (Mezquital), Papiamentu, Paumarí, Piro, Purépecha, Quechua (Imbabura), Romansch, Russian, Samoan, Sango, Santali, Seychelles Creole, Shipibo-Konibo, Spanish, Swahili, Swedish, Tagalog, Turkish, Vietnamese, Warlpiri, Wolof, Worora, Yoruba, Zapotec (Isthmus), Zoque (Copainalá).

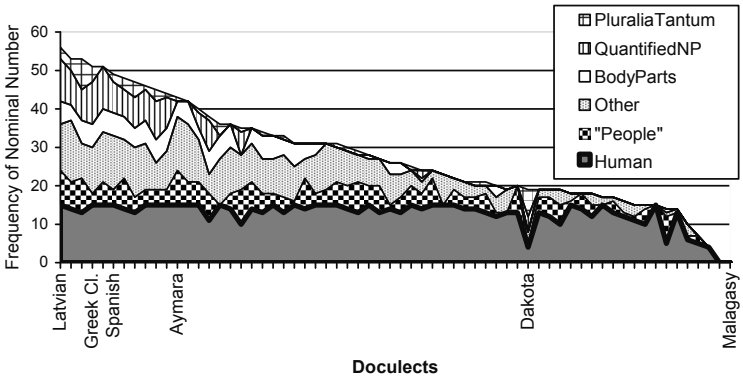


Figure 3. *Semantic profile of nominal number (non-singular) in Mark 8, doculects ordered according to total frequency. The total frequency distribution does not exhibit any extreme poles.*

There is no implicational scale in the exemplar data, but if there is a non-singular marker in any inanimate context there is a good chance that there will be one in many human contexts. The single outlier Dakota, where *-pi* on nouns is actually not nominal plural, but verbal plural extended to nouns of verbal origin, is a clear anomaly in Figure 3. (The small proportion of plural markers in human contexts and many "pluralia tantum" are not in accordance with the profile of languages with similar total frequency.) This is why Dakota should be classified as a language lacking nominal plural along with Malagasy, but is left here in Figure 3 to illustrate an anomaly in the semantic profile.

Interference in translation is certainly an issue (e.g., King James English *heavens* Mark 1:10 via Greek from the Hebrew plurale tantum), but most source languages are Indo-European and Semitic with a high level of frequency of nominal number. Thus, if there is interference, it generally goes in the same direction. A known case of strong interference is Aymara: "En general, el aymara misionero y patrón marcan el plural tanto en los nombres como en los verbos con mucha frequencia y regularidad, en contraste con el aymara descrito en los otros capítulos de este libro" (Briggs 1993: 382). However, Aymara is still lower than Spanish on the frequency scale and has a semantic profile of a language corresponding to a lower total frequency level (no quantified NPs marked for plural). Thus, even in cases of strong interference, languages retain some aspects of their original profile.

2.5. *Matching the constituent order typologies*

In Figures 4 and 5, selections of languages of the two *WALS* typologies are matched with frequency counts from Mark. For constituent order, Figure 4 shows a very good match even though the domains surveyed are not exactly the same (object vs. local phrase). To compare the two classifications I take a 118 language sample (the 100 doculects in Figure 1 plus 18 additional languages with a strong Eurasian bias) and match it with the languages in Dryer 2005d or, if there is no exact match, with closely related languages from Dryer's typology (e.g., Livonian with Estonian). Dryer's sample is so large (1,370 languages) that matching is no major problem. The set of languages compared consists of 108 languages which are plotted in Figure 4. The x-axis is Dryer's discrete typology; the y-axis is the VL ratio scale. All "VO" languages are in a 60–100 % VL ratio range. This is a complete match. All but ten "OV" languages are in a 0–40 % VL ratio range which is a very good match. The exceptional type "No dominant order" has no clear equivalent on the frequency scale. It might have been expected that the languages cluster around 50 %, but they do not since almost all languages have an order preference, even those with "no dominant order". If we consider the languages in 10 % intervals from the poles to be mismatches, two more languages, Motuna and Nunggubuyu, have be added to the set of mismatches which then amounts to 12 of 108 (11 %). Mismatches are due to three factors which partly conspire: (i) differences between the VO and VL typologies; (ii) non-representativity of the translation (VL drift due to interference from VL source languages); and (iii) the difficulty to determine dominant order in languages without rigid order:

(i): Bambara, Bribri, Ngäbere (the latter two Chibchan) are OVL languages and Moru has consistent VL, but the OV/VO order alternation is grammatically determined.

(i) and (ii): Tobelo has 0.96 VL ratio in Mark and 0.56 VL ratio in a counting in the original texts given in Holton (2003: 71–83). Dryer's source Holton (2003: 54–55) says: "Tobelo word order is generally SOV, though other patterns are possible [...] Oblique arguments may [...] occur either before or after the verb". Similarly, in Piro (now called Yine) SOV order is the rule according to Matteson (1965: 38). The grammar is not explicit about VL/LV order, but the text specimens show that there is no consistency in order in original texts, with VL and LV about equally frequent. The parallel text shows a clear preference for VL (0.78), probably due to the influence of Spanish.[3]

3. Interestingly, the grammar and the translation of Mark are from the same author.

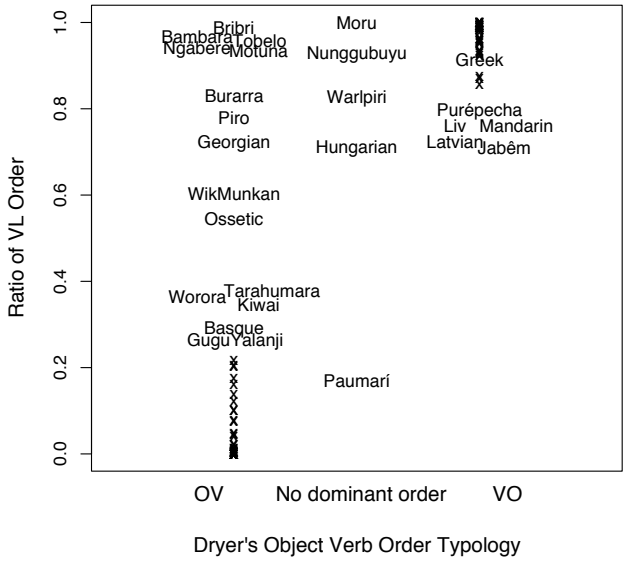


Figure 4. *Matches and mismatches between object verb order type (Dryer 2005d) and VL ratio in motion events in Mark (languages too tight to be labeled rendered by "x")*

(ii): In Motuna there is a strong preference for VL in Mark (0.94) although the Motuna texts edited by Ōnishi (2003) show a clear preference for LV order (VL ratio of 0.15 in 53 examples).

(iii) and (ii): Georgian and Ossetic are known for their flexible word order. It is possible that OV and LV prevail at least in some texts, but the difficulty of classifying these languages is certainly not caused only by interference. The same holds probably for the three Australian languages Burarra, Nunggubuyu, and Wik Mungkan, none of which has rigid order.

As expected, the "translation universals" leave some traces in the parallel text data, there is some interference (Motuna, and to a lesser extent Piro) – in general there is a slight overall VL drift – and there is some "grammatical conventionalism", viz., more rigid order than in original texts. But there is no language with rigid order whose typology is completely distorted.

The discrete *WALS* classification also has its disadvantages. Even though there are no errors strictly speaking in the 118 of 1,370 languages selectively evaluated (8.6 %), the small portion of languages without clearly dominant order are not represented accurately, which is due solely to the mode of the data, the three discrete classes. "No dominant order" is no coherent group, it is no type of the typology, but only a set of different exceptions to the typology and

there is no sharp difference between "No dominant order" and "OV" on the one hand and "No dominant order" and "VO" on the other hand.

2.6. *Matching the nominal number typologies*

The match is not equally good in the case of nominal number and there is no hope of seeing any tendency toward a full match in this case even after removing some exceptions due to inaccurate translation. In the case of nominal number we can compare the intersection of Haspelmath's 290 languages with the 82 languages of the Mark dataset, which gives us 35 languages (12 % of Haspelmath's sample). This might be too little to prove an exact match, but it is enough to show the many mismatches displayed in Figure 5. Rather than explaining away mismatches, let us point out the rather limited matching potential of the typologies. As expected, the "obligatory" sets have higher frequency peaks than the "optional" sets and the "optional" sets have higher peaks than the "none" set. Put differently, there is a good chance that languages with high frequency of nominal number are classified as "obligatory", but not all languages classified as "obligatory" have a high frequency. There is a good chance that languages classified as "optional" have a moderate or low frequency and that languages classified as "no" have a very low frequency at least in the inanimate domain. However, the terms "none", "obligatory", and "optional" cannot be taken seriously in their literal meaning. They are nothing more than conveniences of description in a strong data reduction typology and each of them has a particular range when compared to a related lower data reduction typology. Surprisingly, the typologies do not even correspond for "none".[4]

Haspelmath (2005) is well aware of the high degree of data reduction that his typology implies. As a necessary consequence of using reference grammars as

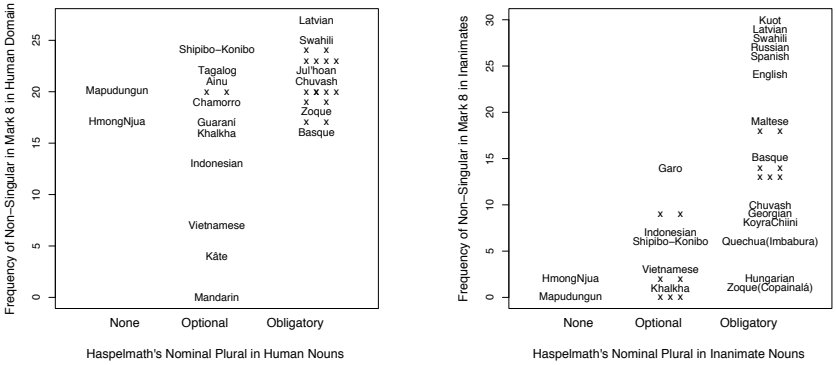
4. The only language of the thirty-five lacking nominal plural in Mark 8 is Mandarin, which has a plural marker that happens not to occur. Hmong Njua has a group classifier which is used as a plural word (*cov*) and Mapudungun has a plural word noted already by de la Grasserie (1886/1887: 237; see also Dryer 2005a). Interestingly, in Mapudungun the Bible translation is more conservative than modern varieties. According to Fernando Zúñiga (personal communication), the plural marker *pu* is expanding in use in modern dialects due to Spanish influence to such contexts as *ñi pu pewma* 'my dreams' and *mi pu rakiduam* 'your thoughts', while it was completely or almost completely restricted to inanimate use in older varieties. In contrast to Aymara, the Bible translation reflects the older use here. This example reflects a general problem of areal typology. Languages cannot be simply restitued to the precolonial state. Influences of Spanish in Latin America and Indonesian in languages of Indonesia (the case of word order in Tobelo) are instances of on-going contact-induced change. Such examples become more manifest with lower levels of data reduction, but it cannot be simply assumed that contact-induced change of the colonial period can be cleansed away by massive data reduction.

a data source, the levels of description in the typology must be so general that most grammars can be assumed to note it (Haspelmath 2005: 143):

> So, like animacy, the dimension of obligatoriness consists of more distinctions than are made here, but since many grammars do not say what happens when a noun is combined with a numeral, it was necessary to simplify the picture and loosen the criteria for obligatoriness.

However, Haspelmath must go even a step further and interpret the lack of explicit information in a grammar as a default value: "Plural marking is classified as optional when it is explicitly described as being optional", put differently, if the grammar does not say anything about use, it is obligatory. This strategy of interpreting no information given in the source as information is an obvious source of misclassifications. The result is that the type "obligatory" covers a broader range of frequency levels than "optional" and "none", which is exactly what is found when the discrete types are matched with the continuous text-based dataset.

As shown here, some types in *WALS*, such as Haspelmath's "obligatory" nominal number in inanimate nouns, cover a very broad frequency range. In this case this is not unexpected if one reads the chapter related to the map carefully. However, serious problems will arise as soon as the database is further processed statistically and occurrence of nominal number is compared with word order in terms of areality and stability. The two typologies are not commensurable because their discrete types in the strong data reduction mode are not equally sharply distinct.



(a) Nominal non-singular in the human domain

(b) Nominal non-singular in the inanimate domain

Figure 5. *Matching ranges in occurrence of non-singular typologies ("x" is used where there are too many languages in the same region)*

2.7. *Conclusions*

The distinction of any pair of types in a discrete typology is the more accurate the more languages in the two sets clearly differ in the intended dimension. Selective evaluation on the basis of different material suggests that the distinction between the types "OV" and "VO" in Dryer 2005d is very accurate, while the other distinctions between the three types in the two typologies surveyed are accurate to a much lower extent. The two types of the accurate distinction happen to be poles of a strongly bimodal frequency distribution. Many other *WALS* typologies are based on functional domains with salient bimodal distributions which remain to be explored in more detail in studies based on additional datasets. However, it is likely that discrete typologies of continuous properties are equally accurate to the extent that they happen to go together with bimodal distributions, and that *WALS* authors intuitively preferred typologies with extreme distributions that lend themselves most easily to extreme data reduction. *WALS* is thus likely to have a strong bimodal distribution bias in the features surveyed. The consequences are that, while many languages are well classified, those not located at the extreme poles are not and that *WALS* suggests a parametric organization of grammar without adducing any evidence for it. This is further discussed in the next section.

3. Discrete features obtained by data reduction do not support parametric organization of grammar

In the generative tradition and beyond, many linguists believe that linguistic categories are universal and that grammar is essentially organized parametrically in the form of discrete features and rules. One motivation to compile a typological database with discrete features is therefore to directly catch parameters of universal grammar as explicitly intended by Guardiano & Longobardi (2005), surveying thirty parameters of the syntax of noun phrases in fifteen languages (fourteen of them Indo-European), and Longobardi & Guardiano (forthcoming), surveying fifty-one binary parameters in twenty-six languages. According to Guardiano & Longobardi (2005), parameters have the following characteristics: (i) discreteness (no continuum), (ii) finiteness (finite number of parameters), and (iii) limitedness with respect to the differences appearing on the surface. For instance, their Feature 3 "gramm. def in DP" is not much different from the feature surveyed on *WALS* Map 37 "Definite Articles" in Dryer 2005b except in terminology and except that the *WALS* map covers many more languages and some more information than simply the presence of definiteness markers in noun phrases.

Interestingly, at least some authors of *WALS* have a diametrically opposite point of view. Haspelmath argues explicitly against parametric explanations (forthcoming a) and holds that "typology must be (and usually is) based on

a special set of COMPARATIVE CONCEPTS that are specifically created by typologists for the purposes of comparison" (forthcoming b). He claims that "those large-scale crosslinguistic studies that were carried out by generativists [...] have in practice used comparative concepts". What I would like to add to this is that typologists create comparative concepts with particular degrees of granularity not given *a priori*. Therefore, discrete features obtained by data reduction do not support the parametric organization of grammar.

Obviously, constituent order in a language is non-parametric, because it is not underlyingly discrete – like the sex of a person is discrete while water levels are not underlyingly discrete even if measured simply in terms of "low" and "high". It may be objected that the parameter is only relevant for a majority of "rigid" order languages and that "free" order languages lack that parameter setting. However, this is not consistent with the finding that nearly all languages exhibit order preferences, even if lacking rigid order.

Of course, there must be some device that causes word order to preferentially assume extreme values, such as the water level in a sluice which is the same or nearly the same as that of the river or canal on either side of the sluice most of the time and which in a sluice at work is not diachronically stable at an intermediate level. However, there is no parametrical organization in sluices that prevents them from having intermediate water levels.

This does not directly imply anything about the general organization of grammar. There are many options and grammar is likely to be organized partly in discrete and partly in continuous terms. One solution is to assume that constituent order is an epiphenomenon of discrete parameters or rules interacting with discourse. Another solution would be to assume that there are no strict rules at all. All instances of verb-object constructions are closely related semantically. We know very well from the semantic map approach and related approaches that related meanings tend to have similar form (Haiman 1985: 19, Haspelmath 2003: 230, Wälchli & Cysouw forthcoming). If all exemplars of verb-object constructions are closely related semantically, they will tend to exhibit similar form, which causes a strong tendency to use the same order of elements in all instances (implemented in discourse by priming). This tendency will be the stronger the less there is deviation, but it is at work even in languages lacking rigid order.

However, the study of typological features beyond the high data reduction mode is indispensable for conveying more evidence for a more differentiated theoretical discussion of the organization of grammar.

4. Conclusion

An anonymous reviewer summarizes my paper as follows: "What is interesting here is the suggestion that some typological features are more amenable

to study in a 'data reduction' mode and others are not. Presumably the latter can be studied better by comparing texts rather than information gleaned from reference grammars. This is the point that should be emphasized in this paper." However, it is not always possible to know in advance (other than intuitively) which typologies are more amenable to study in a "data reduction" mode. This can be shown only by considering distributions in texts. Thus, studies comparing texts are needed not only for typologies where they are the only option, but also for testing where they are the better or a complementary option. As typology evolves, it must make efforts to render its intuitions more explicit. This has been done already in the literature about language sampling, but rarely elsewhere.

Received: 25 July 2008
Revised: 5 January 2009

Universität Bern

Correspondence address: Institut für Sprachwissenschaft, Universität Bern, 3000 Bern 9, Switzerland; e-mail: bernhard.waelchli@isw.unibe.ch

Acknowledgements: I would like to thank to Joan Bybee, Martin Haspelmath, and two anonymous reviewers for their useful comments and to the Swiss National Science Foundation (PP001-114840) for funding.

References

Briggs, Lucy Therina. 1993. *El idioma aymara: Variantes regionales y sociales*. La Paz: Ediciones ILCA.

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

Cysouw, Michael. 2002. Interpreting typological clusters. *Linguistic Typology* 6. 69–93.

Cysouw, Michael & Bernhard Wälchli (eds.). 2007. Parallel texts: Using translational equivalents in linguistic typology. Theme issue in *Sprachtypologie & Universalienforschung* 60(2). 95–181.

Dryer, Matthew S. 2005a. Coding of nominal plurality. In Haspelmath et al. (eds.) 2005, 138–141.

Dryer, Matthew S. 2005b. Definite article. In Haspelmath et al. (eds.) 2005, 154–157.

Dryer, Matthew S. 2005c. Order of subject, object, and verb. In Haspelmath et al. (eds.) 2005, 330–333.

Dryer, Matthew S. 2005d. Order of object and verb. In Haspelmath et al. (eds.) 2005, 338–341.

Dryer, Matthew S. & Orin Gensler. 2005. Order of object, oblique and verb. In Haspelmath et al. (eds.) 2005, 342–345.

Grasserie, Raoul de la. 1886/1887. Etudes de grammaire comparée: De la catégorie du nombre. *Revue de Linguistique* 19. 87–105, 113–146, 232–253, 297–323; 20. 54–67.

Guardiano, Christina & Giuseppe Longobardi. 2005. Parametric comparison and language taxonomy. In Montserrat Batllori, Maria-Lluïsa Hernanz, Carme Picallo & Francesc Roca (eds.), *Grammaticalization and parametric variation*, 149–174. Oxford: Oxford University Press.

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

Halliday, Michael A. K. 1991. Towards probabilistic interpretations. In Eija Ventola (ed.), *Functional and systemic linguistics: Approaches and uses*, 39–61. Berlin: Mouton de Gruyter.

Harrell, Frank E. 2001. *Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis.* New York: Springer.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language,* Vol. 2, 211–242. Mahwah, NJ: Lawrence Erlbaum.

Haspelmath, Martin. 2005. Occurrence of nominal plurality. In Haspelmath et al. (eds.) 2005, 142–145.

Haspelmath, Martin (forthcoming a). Parametric versus functional explanations of syntactic universals. To appear in Theresa Biberauer (ed.), *The limits of syntactic variation*. Amsterdam: Benjamins.

Haspelmath, Martin (forthcoming b). Comparative concepts and descriptive categories in cross-linguistic studies. Draft, July 2008.

Haspelmath, Martin, Matthew Dryer, David Gil & Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Cambridge: Cambridge University Press.

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

King, Alan C. & Begotxu Olaizola Elordi. 1996. *Colloquial Basque: A complete language course*. London: Routledge.

Koptjevskaja-Tamm, Maria & Bernhard Wälchli. 2001. The Circum-Baltic languages: An areal-typological approach. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), *Circum-Baltic languages*, Vol. 2: *Grammar and typology*, 615–761. Amsterdam: Benjamins.

Lewis, Geoffrey L. 1967. *Turkish grammar*. Oxford: Clarendon.

Longobardi, Giuseppe & Christina Guardiano (forthcoming). Evidence for syntax as a signal of historical relatedness. *Lingua*.

Maddieson, Ian. 1984. *Patterns of sounds.* Cambridge: Cambridge University Press.

Maddieson, Ian. 2005. Consonant inventories. In Haspelmath et al. (eds.) 2005, 10–13.

Matteson, Esther. 1965. *The Piro (Arawakan) language*. Berkeley: University of California Press.

Mauranen, Anna & Pekka Kujamäki (eds.). 2004. *Translation universals: Do they exist?* Amsterdam: Benjamins.

Ōnishi, Masayuki, with Dora Leslie & Therese Minitong Kemelfield. 2003. *Motuna texts* (Endangered Languages of the Pacific Rim A1-006). Kyōto: Nakanishi.

Wälchli, Bernhard (in preparation). Motion events in parallel texts.

Wälchli, Bernhard & Michael Cysouw (forthcoming). Toward a semantic map of motion verbs: Explorative statistical methods applied to a cross-linguistic collection of contextually-embedded exemplars. *Linguistics*.