Does sampling matter? A test in replicability, concerning numerals

by THOMAS WIDMANN and PETER BAKKER

1. Introduction

To what extent do we reach different conclusions in typology when we use different samples of languages? This is a question that has often been raised by typologists, but seldom has an attempt been made to answer it. The current paper is a contribution to this problem: Do we obtain significantly different results depending on the samples used? To what extent do size and sampling method matter? These questions involve the reproducibility of results obtained.

Let us first compare sampling of languages for typological purposes with a familiar type of sampling: surveys of voting behavior before political elections. Often several polling institutions, using sampling methods that involve a representative part of the whole population of voters, present their predictions for upcoming elections. Typically, there are differences between the various polls, and after the elections it will become clear whose predictions were closest to the actual electoral results. In other words, one finds out, after the elections, which of the polling institutions used the best sample and extrapolation techniques, and thus the best approximation of reality.

In typology, we have never been able to find out whether a language sample actually was representative for what we hoped to measure, because there are no events comparable with elections where one can observe which polling institution (read: which sample) gave the best predictions (read: representativity of the population), and one has had to assume there that the validity of a study was not compromised by the sample used. The current study could be seen as a method for an evaluation of typological samples and also as a method of replication. We selected a number of existing samples and measured them against what could metaphorically be seen as the result of elections, viz., a maximum subset of the known languages of the world. This subset contains more than 4,000 languages, which is more than half of the languages known to exist.

It might be worth stressing the fact that even if we could investigate every single one of the world's known languages, this would still not give any final answers to questions about language universals; what we are focusing on here is trying to harvest the known linguistic variation by looking at a sample. We are also well aware of the fact that there are many gaps, mostly extinct languages and undescribed languages.

Generally, one can attempt to replicate a typological sample study in at least three ways. One can use the same sample with the same research questions (e.g., the position of negative markers). This is perhaps the least satisfying type of replication, and therefore almost never executed, and never in typology. One can also study the same question with a different sample. This is what Haspelmath & Siegmund (2006) did with Greenberg's word order universals, yielding, in their view, very similar results with a number of different samples. Finally, one can use existing samples, but this time for a different research question and evaluate the results. It is the last type of reproducibility that we test here.

We shall use a number of existing typological samples applying them to novel sets of data, and compare the results with other samples, including one huge sample of 61 % of the known languages of the world. One may criticize our use of the word "sample" here, claiming that "subset" is more appropriate. This should be kept in mind when we use "sample" for this 61 % subset.

In other words, our study tests the samples themselves, admittedly in a different context from the one the sample was created for. This is done by evaluating the results of the existing samples against results obtained from a much bigger sample (the "maximum" sample or subset of more than half of the world's languages). The results of the different samples are compared with the results of the maximum sample and evaluated. The research questions used for this test involves the typology of numerals from 1 to 10 – simply because these data are readily available.

Most of the typological samples we have compared can be characterized as variety samples or probability samples: the compilers all had as their goal to obtain a representative sample of the languages of the world by trying to present data from languages spread over all parts of the globe, languages from as many genetic lineages (genera, stocks, families) as possible, avoiding genetic and areal bias. Probability samples are samples in which each language has an equal chance of being selected – in some cases that means from a genetic lineage. Further, one of them is a random sample (see below). Rijkhoff & Bakker (1998, hence RB) describe a method of sampling, and we applied this and ended up with 182 languages. We wrote a program to compute the largest possible variety sample for a given set of languages; given the shortage of Papuan data, the maximum sample size turns out to be 182, since any larger sample would be lacking data from at least one subfamily. For all of the samples, one of the goals was the exploration of the diversity of human languages. Some of these samples were compiled using more or less sophisticated and explicit methods. Other samples could be characterized as convenience samples in that the availability of accessible studies played a much bigger role than in the others (ultimately, all samples may end up with blanks when data appear to be unavailable). The sample sizes varied from 30 (Greenberg) to 443 (Dryer). An overview of the ten samples and brief characterizations can be seen in Table 1. Three of them can be called convenience samples, based on the

Table 1. *The samples investigated (the sample size is the size of our replication, not the original size)*

Name	Description	Our sample size
Maximum	All languages as described	4,375
Random	Every 18th language from an alphabetical list (see text)	392
Greenberg	A convenience sample (Greenberg 1966)	30
Dahl	A large convenience sample (Dahl 1979)	234
Haspelmath	A convenience sample (Haspelmath 1997)	53
Perkins	A so-called hologeistic sample (Perkins 1992)	49
RB	A variety sample (Rijkhoff & Bakker 1998)	182
Stassen	A non-automatic variety sample (Stassen 1997)	409
GRAMCATS	A probability sample (Bybee et al. 1994)	76
Nichols	A variation-maximizing sample (Nichols 1992)	174
Dryer	A genera sample (Dryer 1992)	443
Bybee	Similar to Perkins, but with modifications (Bybee 1985)	50

characterizations by the authors themselves: Greenberg (1966), Dahl (1979), and Haspelmath (1997). Stassen's (1997) sample was apparently not compiled in a pre-determined manner, and he purposely included some closely related languages and several languages from one family, when that family displays particularly interesting phenomena for the research question involved (Stassen 1997: 8, 617: Note 5; see also Rijkhoff & Bakker 1998: 303).

2. Method

We decided to use numerals as a testing ground, because a list of numerals was readily available, Rosenfelder (2001). Compiled over the course of many years, and constantly updated, it can be found on http://www.zompist.com/numbers.shtml. At the time of testing, in October 2001, this website contained the numerals from one to ten in 4,584 languages (in July 2005 over 5,000, and now covering 83 % of the world's languages according to the compiler). The 5,000 figure also covers conlangs, reconstructed languages and dialects, i.e., not always distinct languages in a strict sense of the word.

The reliability must be considered very high. The website is consulted by many, and there are frequent updates. We checked some random languages, and we could not spot any errors. Harald Hammarström (personal communication) points out a number of problems with this dataset. The website reliably copies the (published) sources, but in a number of cases existing numerals are lacking, there are spurious languages, and sometimes borrowed numbers are omitted. We are aware that these data are not perfect, but its availability and extent makes this a convenient source for a study on replication.

Table 2. *Numbers in Paravilhana (Macro-Carib), a language with its "highest" number below 10, and Nek (Central-Western Trans-New Guinea), a language with gaps*

Number	Paravilhana	Nek
1	*teuenjé*	*nogan*
2	*akoünien*	*tipet*
3	*olaulé*	*tipetkutno*
4	*olaulá*	*tipet tipet*
5	*avainjanlö*	*kitombo*
6	*adöu avainjanlö*	*kitombo nogan*
7	*enepù náei*	(not recorded)
8	*olau lei*	(not recorded)
9	(not recorded)	(not recorded)
10	(not recorded)	*kit tombon tombon*

From a European perspective one could wonder how these lower numerals could be typologically interesting since virtually all European languages show discrete numbers from 1 to 10 (called "simplex" systems here, i.e., without mathematical operations at all below 10, such as 7 as $5+2$). In all languages, including European ones, we get mathematical operations on numbers above 10, such as addition ($16 = 6+10$) and multiplication ($60 = 6 \times 10$). Indeed only six percent of Europe's languages in the maximum sample display at least one mathematical operation for numbers 1 to 10. On the other hand, no fewer than 67 % of Papua's languages do so (and 39 % of the maximum sample). In other words, quite a few languages use mathematical operations also for numbers 10 and lower. Furthermore, some languages have their "highest" number (that is, the largest known number) below 10 and some languages display gaps (e.g., they lack a number 9, but do have 10). Note that gaps are most likely omissions due to lack of data rather than being genuine gaps. Examples of the two types just mentioned are the number systems of Paravilhana (Macro-Carib) and Nek (Central-Western Trans-New Guinea) given in Table 2. Harald Hammerström points out that the ultimate source for the Nek numbers is Smith (1988), where it is clear that numbers 7 to 9 do exist, and are regularly formed, hence are not a real gap in the counting system.

Languages do not only make use of addition and multiplication. Subtraction is also found in counting systems 1 to 10, where, e.g., 9 is given as $10-1$. Division has not been identified in the lower numerals, but it does exist in higher numerals (e.g., Breton *hanterkant* 'half hundred, i.e., 50'; compare also English *half a dozen* or Danish *halv snes* 'approximately ten', both outside the normal counting system). There is a fair amount of literature on the typology of numerals. Interested readers can consult Comrie (1997), Greenberg (1978), Gvozdanović (1999), Heine (1997: Ch. 2), Hurford (2003), Seiler (1990), and Stampe (1976), and dozens of additional monographs and hundreds of articles.

We obtained the data from the website and downloaded them for further processing. The 200 or so constructed languages (Esperanto, Novial, Quenya, Klingon, etc.) on the website were removed (reconstructed languages were kept). After this, the languages were matched with the languages listed in the 14th edition of *Ethnologue* (Grimes (ed.) 2001) to obtain their three-letter language code. Languages with the same language code were considered to be dialects of the same language, and the first one of those in the alphabet was systematically selected as the only representative of this dialect cluster. In this way 406 number systems have been disregarded. Thus we were left with a usable sample of 3,969 languages. If there are 7,148 languages in the world, this maximum sample would constitute 56 % of the world's languages (the *Ethnologue* website www.ethnologue.com mentions 6,809 languages, but they do in fact have 7,148 languages listed). When number systems were available for a language not identifiable in the *Ethnologue*, we assigned it an asterisked two-letter code (346 languages in total). Thus we ended up with 4,375 languages, 61 % of the world's known languages.

Note that if we assume that if anything is known about a language, it may be the lower numerals, then this means that we do not even have the most basic information about almost half of the world's languages. On the other hand, Rosenfelder's enterprise does not cover all of the material available in print, and the database is constantly increasing, suggesting that in fact much more is available.

4,375 languages is a huge number, so we automatized procedures as much as possible. We wrote a computer program that should recognize recurrent patterns in the numeral systems. This will lead to errors in two directions. It may include accidental similarities, such as Frisian *tsien* '10' and *ien* '1'. It may also miss similarities because they are hidden by morphophonological processes, like Finnish *kahdeksan* '8' and *yhdeksän* '9' which are cognate with *kaksi* '2' and *yksi* '1' (cf. inflectional stems *kahde-*, *yhde-*). We assume such errors to be evenly distributed, but this is difficult to prove. This obviously produces a bias, as is clear when one compares, e.g., Europe, where all languages have numerals 1 to 10, with South America and Australia, where higher numerals are often lacking, and thus they display no similarities. We have not corrected such errors when we did identify them, in order not to add bias to the sample.

In order to simulate the typological samples, we selected the number systems of the languages used in the samples mentioned above. When we could not find the number systems for a particular language in a sample, we replaced the language with the most closely related one for which we did have information. This was necessary when the language used in a sample could not be identified in the list of numbers, or because it was lacking. These samples were tested and compared along a number of typological properties in the number systems, some of them extremely common, others of them rare.

In addition, we employed a random sample, i.e., a sample of randomly chosen languages, a method not used before in typology – as far as we know. This was a so-called systematic sample, i.e., by choosing every *n*-th language of a list sorted by *Ethnologue* language codes. Many other ordering criteria could have been employed instead so long as all languages would still have had equal chances of being picked; the *Ethnologue* language codes were chosen because they were assumed to be fairly random. Having chosen the 7,148 languages in *Ethnologue* and using Scheaffer et al.'s (1996) formula counting with variance and expected variation, we ended up with choosing every 18th language and thus a sample of 392 languages, so that we end up with roughly 5 % of the maximum sample. Random/systematic samples are a standard technique in statistics. In random samples there is a risk that widespread language families and language types tend to dominate the picture. That is probably one of the reasons this technique has not been used in typology.

3. Results

First we calculated what proportion of the languages in the different samples identified the same number of simplex systems as the maximum sample, i.e., 71 %. Haspelmath, Dahl, Greenberg, and Perkins fare worst, overestimating the actual number (with 91 %, 81 %, 77 %, and 76 %, respectively), whereas Bybee, Nichols, and Dryer yield the lowest numbers (66 %, 60 %, 60 %). RB, Stassen, and GRAMCATS score best (70 %, 70 %, and 71 %), and Random also scores well (68 %). However, if we take into account that European numerals are better documented, and that several languages do not have discrete numerals, the maximum sample percentage may be too high. Numbers higher than 71 also suggest a Eurasian bias.

Further we calculated the number of languages in each sample in which addition is used with 5 as a base (e.g., 7 as $5+2$) for numerals under 10. These are listed in Table 3. Smaller samples and convenience samples tend to score too low, whereas the largest samples tend to be most widely off. Random, RB, and Stassen score best.

We also calculated the distribution of the highest recorded numeral for the samples. For 85 % this was the number 10, or higher. Greenberg, Dahl, and Haspelmath score too high (97 %, 94 %, 96 %). RB, Dryer, and Nichols have the lowest scores (76 %, 78 %, 82 %). All the other samples, including the random one, score between 84 and 86 %, except Stassen with 90 %.

Numbers 1, 7, and 8 are the highest numerals in the maximum sample in less than 0.5 % of cases, and 6 and 9 both in 1% of it. When 1 is the highest number, it means that the language has no counting system. Actually, 7 and 8 do show

Table 3. *Percentage of languages in which the numbers from 6 to 10 are constructed as additions to 5*

Sample	6	7	8	9	10
Maximum	6	9	7	7	1
Random	6	9	7	9	2
Greenberg	3	3	0	0	0
Dahl	2	4	2	2	0
Haspelmath	0	4	2	0	0
Perkins	0	6	2	6	0
RB	8	8	4	5	2
Stassen	7	10	9	8	1
GRAMCATS	11	11	9	13	0
Nichols	6	15	9	10	1
Dryer	8	10	8	10	2
Bybee	2	10	4	8	0

up as highest numbers in three of the samples (RB, Nichols, Dryer), but these got lost in rounding off in the maximum sample. The numbers 2, 3, 4, and 5 appear as the highest numbers in the maximum sample in two, four, three, four languages (respectively) in a seemingly arbitrary distribution. Similar but different percentages show up in the other samples – except for Greenberg, Dahl, and Haspelmath, which score too low.

We also studied two rare phenomena: first, the existence of a composite number 10 (e.g., $5+5$ or $8+2$). This is found in only 2.8 % of the world's languages in the maximum sample. The highest numbers show up in Dryer (6.5 %) and Nichols (5.2 %), and the lowest in Greenberg (0 %), Dahl (1.3 %), Haspelmath (1.9 %), and Perkins (2 %). RB, Stassen, and GRAMCATS again score best with 2.7%, 2.7 %, and 2.6 %.

Second, we studied systems with gaps, which is an even rarer phenomenon. In fact, many of these gaps may be entirely caused by incomplete documentation (see above). Cultures in which certain numerals are taboo words can contain genuine gaps. Whether or not we are measuring a rare or non-existing or artificially introduced phenomenon, this can be taken as a test to what extent rare phenomena can be discovered in the different samples.

A numeral system with a gap is one in which one numeral is lacking, but the series continues after. Only 0.2 % of languages in the maximum sample show gaps, i.e., they have, for example, a number for 8 but not one for 7. If languages have a gap, it is mostly number 6 (0.4 %), decreasing both upwards (7: 0.3 %, 8: 0.2 %, 9: 0 %) and downwards (5, 4, and 3: 0.3 %, and 1 and 2 both 0.2 %). 10 is not included because it is the highest number of the domain and could therefore be lacking because the system would end here. None of the

Table 4. *The number of different numeral systems in each of the samples*

Sample	Different	Total	Percentage
Maximum	652	4375	15
Haspelmath	8	53	15
Dahl	51	234	22
Stassen	94	409	23
Random	95	392	24
Dryer	130	443	29
Greenberg	9	30	30
RB	61	182	34
Perkins	17	49	35
Nichols	65	174	37
Bybee	19	50	38
GRAMCATS	30	76	39

samples would yield this hierarchy from 6 up and down, but Stassen's sample clearly gets closest. On the low side Greenberg, Dahl, Haspelmath, Bybee, and Perkins score worst, whereas the larger samples of Nichols, Dryer, and RB do reveal some of this. GRAMCATS did well in the other tests, but it scored badly here, with a zero score for all numbers except for number 3 (too high a score of 1.3 %). The random sample also scores fairly badly here, finding only gaps for 2, 3, and 4. For rare phenomena such as these, larger samples are (not surprisingly) better than small samples, but it remains a question of chance. Random sampling does not yield representative results here. The hierarchy starting from 6 (if it is really meaningful, taking into account that many gaps are not real gaps in the system but just gaps in the data) would, by the way, only be discovered in this huge maximum sample.

Finally, it may be interesting to know how many different numeral systems were found in the sampled languages, cumulating all the parameters of difference discussed above (gaps, addition operations, highest numbers) plus a few others not reported on here. In the maximum sample no fewer than 652 different systems were identified, giving a percentage of 15 % of the total sample. The results for the samples are given in Table 4. Not all of the other samples yielded higher scores than the maximum sample (Haspelmath had the same percentage of distinct systems) – which means that sampling is actually better for capturing linguistic diversity than using all accessible languages in a census-type survey. The lowest scoring samples are those of Dahl and Haspelmath, and the highest the samples GRAMCATS, Bybee, Nichols, and Perkins.

4. Areal bias

Some samples clearly have an areal bias, most of all the convenience samples (Greenberg, Dahl, and Haspelmath). Areal bias here means a Eurasian bias. To investigate the amount of bias in each sample, we tried to count the proportion of simplex numeral systems on each continent according to each sample, the result of which can be seen in Table 5. The variance compared to the maximum sample has been calculated for each sample, in order to better evaluate the scores. The figures from Greenberg and Haspelmath are not really comparable because they lack figures for one or more areas (indicated with — in Table 5). Variance is listed in the last column: the closer to zero the score is, the closer the sample scores to the maximum sample. As could be expected, the random sample gives results reasonably close to the maximum sample (which ought to be fairly close to the real figures), but Stassen, RB, GRAMCATS, and Nichols actually give better results in this case, and Dahl is just as good as the random sample. For background information, Table 6 lists the number of languages in each area for each sample.

5. Family bias

We also tried to calculate family bias using the same calculation technique for some of the world's major language families or stocks. Some of these groupings (in any case Austro-Asiatic, to some extent also Nilo-Saharan) are areal rather than genetic groupings. The results can be seen in Tables 7 and 8. Here, the difference between the samples is massive. The random sample does best, scoring closest to zero in the variance column in Table 8, and Dahl and Dryer do fairly well. Perkins and Bybee do very badly, on the other hand. Here again sample size matters: the larger the sample, the better the results. We have added a column in Table 8 with language isolates because the samples differ very much in their treatment of these languages. The proportion of isolates is higher in the smaller samples. In Table 8, we have added a row with the proportions found in the *Ethnologue*, since these must be considered the best available values.

6. Conclusions

All samples considered here want to capture diversity, but not all of them succeed to the same extent. The choice of one's sample does matter. Not all of them yield the same results.

Convenience samples clearly fare worst. Size apparently does not really matter a lot here, since Dahl's 234 language sample does not score significantly better than Greenberg's with 30 or Haspelmath's with 53 languages. This contradicts Haspelmath & Siegmund's (2006) conclusion that Greenberg's sample, despite being eurocentric and biased, did capture roughly the right conclusions.

Table 5. *Percentage of areas having simplex number systems according to each sample*

	Papua	Australia	N America	C America	S America	Oceania	Africa	Asia	Europe	Variance
Maximum	33	61	68	70	72	71	73	84	94	—
Random	31	40	88	42	72	83	71	87	94	.42
Greenberg	—	100	—	0	67	100	86	78	86	(.87)
Dahl	0	50	68	75	73	50	64	85	97	.42
Haspelmath	0	—	50	100	100	100	100	81	100	(.69)
Perkins	50	0	73	75	75	100	83	81	100	.71
RB	50	71	70	78	67	75	65	86	100	.25
Stassen	32	40	70	63	71	76	72	80	96	.23
GRAMCATS	25	67	56	75	75	100	87	82	100	.36
Nichols	26	41	56	73	67	100	74	83	100	.38
Dryer	39	21	53	71	64	80	74	74	100	.46
Bybee	50	0	64	60	67	100	50	75	100	.75

Table 6. *The number of languages in each area for each sample*

	Papua	Australia	N America	C America	S America	Oceania	Africa	Asia	Europe
Maximum	529	137	253	120	304	224	1188	1105	169
Random	68	15	8	24	25	18	99	119	16
Greenberg	0	1	0	2	3	1	7	9	7
Dahl	2	6	25	12	11	4	28	73	73
Haspelmath	1	0	2	1	2	1	4	16	26
Perkins	2	1	11	4	4	2	6	16	2
RB	26	7	23	9	36	4	34	35	8
Stassen	37	20	46	24	28	17	88	92	46
GRAMCATS	12	6	9	4	4	3	15	17	5
Nichols	27	17	41	11	15	4	19	24	7
Dryer	87	24	58	28	80	5	68	78	15
Bybee	2	1	11	5	3	2	6	16	2

Table 7. *Percentage of language families having simplex number systems according to each sample*

	Niger-Congo	Austronesian	Trans-New Guinea	Indo-European	Afro-Asiatic	Sino-Tibetan	Australian	Nilo-Saharan	Oto-Manguean	Austro-Asiatic	Sepik-Ramu	Variance
Maximum	69	68	31	91	86	87	60	70	73	78	62	—
Random	70	78	23	93	75	88	36	56	56	80	71	.39
Greenberg	67	100	—	83	100	0	100	100	0	—	—	(1.29)
Dahl	58	75	0	90	100	100	50	50	100	67	—	(.54)
Haspelmath	100	100	0	100	100	100	—	—	—	—	—	(.58)
Perkins	0	100	0	100	100	100	0	100	0	67	100	1.36
RB	57	81	42	100	71	100	71	80	100	33	100	.73
Stassen	63	69	37	88	93	87	40	62	83	67	0	.69
GRAMCATS	67	67	20	100	100	60	67	100	100	100	0	.84
Nichols	50	83	25	100	100	100	41	67	50	0	20	.99
Dryer	68	100	22	100	75	89	21	72	71	71	47	.56
Bybee	0	100	0	100	0	100	0	67	0	50	100	1.60

Table 8. *The proportion of families in the various samples*

	Niger-Congo	Austronesian	Trans-New Guinea	Indo-European	Afro-Asiatic	Sino-Tibetan	Australian	Nilo-Saharan	Oto-Manguean	Austro-Asiatic	Language isolates
Ethnologue	21	18	8	6	5	5	4	3	2	2	0
Maximum	13	14	2	4	5	5	2	2	0	2	0
Random	12	13	3	7	4	8	1	1	1	2	—
Greenberg	7	7	—	17	7	0	3	10	0	—	3
Dahl	5	3	0	16	3	2	1	0	0	2	1
Haspelmath	6	6	0	40	8	2	—	—	—	—	0
Perkins	0	2	0	2	2	2	0	6	0	8	6
RB	7	9	3	3	3	2	3	2	1	1	1
Stassen	6	8	2	7	6	3	2	3	1	2	1
GRAMCATS	5	8	1	7	8	4	5	3	1	4	0
Nichols	2	3	2	3	2	1	4	2	1	0	3
Dryer	4	3	2	2	2	4	1	4	1	2	3
Bybee	0	2	0	2	0	2	0	4	0	6	8

Greenberg's sample did not score well in our tests precisely because of this Eurasian bias, compounded by the small size. Correlations between different phenomena may be captured even in a bad sample (cf. Haspelmath & Siegmund 2006).

Stratified samples are better than convenience samples. The size of the other, more methodically structured, samples, is apparently not always very significant. Sometimes the smaller samples fare better, sometimes the larger ones. For capturing diversity, some stratification method is more important than the size of the sample. Only for rare phenomena, the larger samples are better at finding patterns, for obvious reasons. However, in a very large sample such as the maximum sample, some phenomena may get buried in the statistic generalizations.

The systematic random sample fares fairly well, suggesting possibilities for this technique in the future. A disadvantage for data other than numerals may be unavailability or inaccessibility of data.

This replicability test shows that one's sampling technique does matter. However, more relevant to the replicability issue is the researcher's interpretation of how one categorizes phenomena. This problem did not show up here, since we took the numerals for granted, but it would be a problem, for instance, when languages use different numbers when speakers count, compared to attributive use ("three stones"), or different numerals for animates and inanimates, and the like. This is probably a trivial problem with numerals, but it would be much more problematic with concepts such as "subject", "adjective", and the like. Not all languages may have such categories, and imposing such categories on languages that do not have them will reflect different interpretations and different amounts of eurocentricity, and hence different results.

Received: 2 December 2004
Revised: 3 August 2005

Collins Dictionaries
Århus Universitet

Correspondence addresses: (Widmann) Collins Dictionaries, Westerhill Road, Bishopbriggs, Glasgow G64 2QT, United Kingdom; e-mail: twid@bibulus.org; (Bakker) Afdeling for Lingvistik, Århus Universitet, Bygning 1410, Ndr. Ringgade, 8000 Århus C, Denmark; e-mail: linpb@hum.au.dk

Acknowledgments: We are grateful to Hilary Chappell, Jan Rijkhoff, Martin Haspelmath, and three anonymous *LT* referees for their comments, and also to Harald Hammarström, who pointed out a number of flaws in an earlier version that we hope to have corrected. None of them are responsible for remaining shortcomings. Peter Bakker als expresses his gratitude to RCLT, La Trobe University, wher this article was revised.

References

Bell, Alan (1978). Language samples. In Joseph H. Greenberg, Charles A. Ferguson, & Edith A. Moravcsik (eds.), *Universals of Human Languages*, Volume 1: *Method & Theory*, 123–156. Stanford: Stanford University Press.

Bybee, Joan L. (1985). *Morphology*. Amsterdam: Benjamins.

Bybee, Joan L., Revere D. Perkins, & William Pagliuca (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Comrie, Bernard (1997). Some problems in the theory and typology of numeral systems. In Bohumil Palek (ed.), *Typology – Prototypes, Item Orderings and Universals: Proceedings of LP '96*, 41–56. Prague: Charles University Press.

Dahl, Östen (1979). Typology of sentence negation. *Linguistics* 17: 79–106.

Dixon, R. M. W. (1997). *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Dryer, Matthew S. (1992). The Greenbergian word order correlations. *Language* 68: 81–138.

Greenberg, Joseph H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of Language*, 73–113. Cambridge, Mass.: MIT Press.

Grimes, Barbara F. (ed.) (2001). *Ethnologue*. 14th edition. Dallas: SIL International. http://www.ethnologue.com

Gvozdanović, Jadranka (ed.) (1999). *Numeral Types and Changes Worldwide*. Berlin: Mouton de Gruyter.

Haspelmath, Martin (1997). *From Space to Time*. München: Lincom Europa.

Haspelmath, Martin & Sven Siegmund (2006). Simulating the replication of some of Greenberg's word order generalizations. *Linguistic Typology* 10: 74–82.

Heine, Bernd (1997). *Cognitive Foundations of Grammar*. New York: Oxford University Press.

Hurford, James (2003). The interaction between numerals and nouns. In Frans Plank (ed.), *Noun Phrase Structure in the Languages of Europe*, 561–620. Berlin: Mouton de Gruyter.

Nichols, Johanna (1992). *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Perkins, Revere D. (1992). *Deixis, Grammar, and Culture*. Amsterdam: Benjamins.

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

Rosenfelder, Mark (2001). Numbers in 4584 languages. http://www.zompist.com/numbers.shtml

Scheaffer, Richard L., William Mendenhall III, & Lyman Ott (1996). *Elementary Survey Sampling*. 5th edition. Belmont: Duxbury.

Seiler, Hansjakob (1990). A dimensional view of numeral systems. In William Croft, Keith Denning, & Suzanne Kemmer (eds.), *Studies in Typology and Diachrony*, 187–208. Amsterdam: Benjamins.

Smith , G. P. S. (1988). Morobe counting systems. In *Papers in New Guinea Linguistics* (Pacific Linguistics, A-76), 1–132. Canberra: Australian National University.

Stampe, David (1976). Cardinal number systems. *Chicago Linguistic Society* 12: 594–609.

Stassen, Leon (1997). *Intransitive Predication*. Oxford: Clarendon.

Widmann, Thomas Martin (2001). Language sampling for typological studies. Unpublished Master's thesis, Aarhus Universitet. http://www.twid.bibulus.org/speciale.pdf (main text) and http://www.twid.bibulus.org/appendix.pdf (appendix)