Matti Miestamo*, Dik Bakker and Antti Arppe

Sampling for variety

DOI 10.1515/lingty-2016-0006
Received June 16, 2013; revised December 30, 2015

Abstract: Variety sampling aims at capturing as much of the world's linguistic variety as possible. The article discusses and compares two sampling methods designed for variety sampling: the Diversity Value method, in which sample languages are picked according to the diversity found in family trees, and the Genus-Macroarea method, in which genealogical stratification is primarily based on genera and areal stratification pays attention to the proportional representation of the genealogical diversity of macroareas. The pros and cons of the methods are discussed, some additional features are introduced to the Genus-Macroarea method, and the ability of both methods to capture crosslinguistic variety is tested with computerized simulations drawing on data in *The world atlas of language structures* database.

Keywords: genealogical classification, genus, macroarea, methodology, sampling, variety sampling

1 Introduction

Linguistic typology is a branch of theoretical linguistics that bases its understanding of the nature of human language on empirical research of crosslinguistic variation. Language sampling is an integral part of the methodology of crosslinguistic investigations, and the design of samples is one of the central methodological questions in language typology. Different research questions require different types of samples. Rijkhoff et al. (1993: 171) make a distinction between two basic types of typological sampling: probability and variety sampling. PROBABILITY SAMPLES, on the one hand, are meant to be used for the statistical testing of tendencies and correlations, which makes the requirement of the independence of the sampled units important: only on the basis of samples

***Corresponding author: Matti Miestamo,** Yleinen kielitiede, PL 24, 00014 Helsingin yliopisto, Finland, E-mail: matti.miestamo@helsinki.fi
Dik Bakker, Capaciteitsgroep Taalwetenschap, Universiteit van Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands, E-mail: d.bakker@uva.nl
Antti Arppe, Department of Linguistics, 2-40 Assiniboia Hall, University of Alberta, Edmonton, Alberta, Canada T6G 2E7, E-mail: arppe@ualberta.ca

consisting of independent units can one draw valid statistical generalizations. The goal of VARIETY SAMPLES, on the other hand, is to display as much variety as possible in the linguistic realizations of the phenomena under investigation and to reveal even the rarest strategies or types of expression in the domain explored. This is especially important for the universal validity of grammatical theory: a general theory of grammar should be able to account for all attested types of linguistic structures and phenomena.

Several methods whose goals conform to those of probability sampling have been proposed in the literature (e.g., Dryer 1989, see Section 2.2 for more). As for variety sampling, although sampling methodology has been discussed rather extensively in the literature, only two of the systematically codified methods are designed specifically for variety sampling: the Diversity Value (DV) method by Rijkhoff et al. (1993) and Rijkhoff & Bakker (1998), and the Genus-Macroarea (GM) method introduced by Miestamo (2003) and elaborated in Miestamo (2005). This article focuses on variety sampling: it discusses the two methods and then tests how they fare in achieving the goals of variety sampling.

The article is organized as follows. Section 2.1 addresses central issues in typological sampling, especially from the point of view of variety sampling. An overview of sampling methods proposed in earlier literature is given in Section 2.2. Section 3 gives a short introduction to the DV method. The treatment of the GM method in Section 4 is more extensive, since this method has not been very prominently discussed in the literature and some new features also need to be introduced to the method. Section 5 compares the two methods using a computerized test on data in *The world atlas of language structures* (*WALS*) database (Haspelmath et al. (eds.) 2005; Dryer & Haspelmath (eds.) 2013) Section 6 discusses some further issues and future prospects, and concludes the article.

2 Background

2.1 Variety sampling

The goal of variety sampling is to maximize the amount of variation in the language data. Obviously, all the variety that existing languages exhibit can be covered by including every one of the world's approximately 7,000 languages in the study. However, this is not possible for a number of reasons: not nearly all languages have been sufficiently described to allow for a proper analysis of the phenomenon under investigation; even if sources exist, they might not be accessible to the researcher; and even if one had access to adequate sources

for all 7,000 languages, it would (in the case of most research questions) be too time-consuming to include them all. Variety sampling is therefore needed. In very general terms, we can well presume that genealogical, areal, cultural, typological, etc. connections between languages increase the possibility that they are similar to each other. The more diverse a sample is in these respects, the more linguistic variety it may be expected to show in the domain under study. To achieve its goal of reaching maximal variety, a variety sample should therefore represent all the world's linguistic groupings – areal, genealogical, and other – as well as possible. The better all groupings are represented, the bigger the chances that all the relevant variety is found in the sample.

Probability samples have different goals: when one is interested in cross-linguistic frequencies of features, correlations between them, or other statistical measures, it becomes crucial that the sample has as few biases as possible that could distort the numbers. Crosslinguistic tendencies can be due to historical or areal connections, and these should be controlled for if one wants to capture universals and/or factors (functional/cognitive/social) that shape linguistic structure. It is therefore important that the sampled languages are as independent of each other as possible in terms of genealogy and areal contacts. When sampling languages, full independence is very difficult to achieve. Probability samples have to find a balance between representativeness and independence. Note, however, that some recent developments in testing universals, e.g., the method for genealogical control by Bickel (2008), actually presuppose that samples include related languages from all relevant genealogical groupings, and the focus in testing universals is being shifted towards observing diachronic processes rather than synchronic distributions (see Maslova 2000; Levinson et al. 2011; Cysouw 2011).

In his discussion of language sampling, Bell (1978: 126) introduces the terms "universe", "frame", and "sample", and he defines them as follows: (i) the UNIVERSE is the class of objects which is the object of investigation; (ii) the FRAME is the means of access to the universe; (iii) the SAMPLE is the collection of objects that are observed. In a typological study that aims at making generalizations about natural languages, defined as languages that are or have been used as the native language of a group of language users, the universe is the set of all natural languages, whereas the frame consists of the languages for which one can find data (either from published or unpublished written sources or by asking experts or native-speaker informants). In such a study, the sample should be representative of all natural languages.

Although areal questions are becoming more and more central in the field of typology (see Bickel 2007), it remains the case that a lot of typological work is equally interested in the theoretical limits of crosslinguistic variation, i.e., in the

question of what is a possible natural language, and in crosslinguistic frequencies and correlations motivated by linguistic principles (rather than by extralinguistic historical factors). In this type of work, the universe consists not only of all existing or historically attested natural languages, but also includes all languages that have existed but have changed and become new languages or disappeared without leaving a trace.[1] Furthermore, it is plausible to assume that even all the languages that have existed throughout human history have not exhibited every linguistically possible feature. Therefore, in principle, the universe in variety sampling is the set of all possible languages – past, present, or future. In practice, inferences have to be based on attested languages, which are only partially representative of the possibilities and tendencies of human language. Given that the size and distribution of different linguistic groupings in the world is determined by non-linguistic factors, it is unlikely that the frame is representative of the universe of all possible languages. This may be less of a problem when dealing with features that tend to change fast than with more stable features.[2] In any case, adequate sampling techniques help to bridge the gap between the frame and the universe by reducing the effect that extralinguistic historical factors might have on the selection of sample languages. This means, for example, that the chances of a genealogical grouping being represented in the sample should not be dictated merely by the size of the grouping – a factor that is clearly due to extralinguistic historical causes.

The number of languages to be included in the sample depends on several issues. For probability samples, sample size is closely linked to the question of the balance between representativeness and independence. The larger the sample, the more genealogical or areal connections there tend to be between the sample languages. Even with relatively small sample sizes, it is impossible to include only languages that are completely independent of each other in these

1 Bakker (2011: 101) reasons that if it is assumed that human language has existed in its present form for 40,000 years (which is a rather cautious assumption), that the number of languages spoken in the world has been around 6,000 throughout that time, and that a language changes at such a rate that it becomes a new language every 1,000 years, the number of languages that have ceased to be spoken is around 233,000. This is naturally a very rough estimation, but it nicely illustrates the point that the number of languages that have been spoken during human history must be much higher than the number of attested languages.

2 Maslova (2000: 324–325) proposes that distributions of linguistic features were affected by extralinguistic factors (birth/death processes) only at a very distant stage when the language population was small, but that at more recent stages, at least for the whole period within the reach of the comparative method, distributions have been, to any significant degree, affected by type shift processes only. Be this as it may, it should be noted that type shift processes can also be affected by extralinguistic factors such as language contact, spreading phenomena present in languages of powerful communities.

respects (see Dryer 1989). Any sample has to compromise independence to some extent.[3] Perkins (1989: 312) recommends "using around a hundred languages for most linguistic samples to balance the requirements for representativeness and independence in samples".

For variety samples, the problem is not equally important, however. For these, the mutual dependence between the sample languages is not a problem in any direct sense, and the more languages a variety sample contains, the better it may capture the crosslinguistic variety of the phenomenon under study. Himmelmann (2000: 9–10) even argues that paying attention to sampling is not important when one's primary interests are not in crosslinguistic frequencies. However, if two samples of the same size are compared, the one with more genealogical and areal diversity, i.e., less mutual dependencies among the languages, will likely show more linguistic variety – maximizing the genealogical and areal distance between individual sample languages increases the variety covered by the sample. Independence is therefore an important goal in variety sampling as well.

Himmelmann (2000: 9–10) further claims that at the stage of building hypotheses it is counterproductive to use extensive samples. It is true that hypotheses are often built on the basis of a smaller number of languages, but it should be stressed that even such small pilot samples should be areally and genealogically balanced. This will reduce the danger of untested, false hypotheses starting a life of their own as "scientific results".[4] Sampling methods in which languages are selected from all of the more or less independent genealogical and/or areal groupings are often thought to be able to produce samples that can be representative even with rather small sample sizes (Bybee et al. 1994; Rijkhoff & Bakker 1998). However, as Bell (1978: 142–144), Stassen (1997: 8), and Stolz & Gugeler (2000: 54–55), among others, have pointed out, to give a better picture of the entire range of crosslinguistic variation, variety samples should be more extensive. An extensive sample makes it more probable that no linguistic features, not even the rarest ones, go unnoticed. Smaller samples can reveal what is common and give an idea about how common it is, but the coverage of rarer features is quite random in small samples.

3 As Perkins (2000: 351) notes, independence is never absolute and always a matter of degree. There are statistical techniques to deal with hypothesized biases.

4 Note also that since crosslinguistic frequencies are easily available in the *WALS* database, but many of the underlying studies are not based on adequate samples (cf. also Hammarström 2009), there is a potential risk of these frequencies triggering wrong conclusions. This could have been avoided if the studies had followed adequate sampling methods from the beginning.

The size of a sample is naturally also dependent on the nature of the research question. Note that both Perkins (1989) and Rijkhoff & Bakker (1998) introduce methods of calculating the ideal sample size for a given object of study according to how many possible or expected variables there are, or how these variables are estimated to be distributed; such a measure is naturally only usable in cases where the possible variables are known beforehand.

Although representativeness and independence are required to different degrees depending on whether one is dealing with probability or variety samples, they remain important goals for both types of sampling. Whereas sample size is important for representativeness but potentially harmful for independence, stratification is needed to achieve both of these goals. Stratification means that the sampling universe is divided into relevant subgroups, e.g., genealogical or areal, and instead of picking languages randomly from an unstructured list of the world's languages, random selection is made inside each subgroup so that the representation of the subgroups is not biased in the sample. Different sorts of bias may be harmful, depending on the research questions one has posed. The most general and obvious sources of bias are genealogical and areal connections between the sample languages. Genealogy and geography are thus the most important and the most common bases of stratification. Most samples try to make sure that genealogical and areal groupings of languages are represented in a balanced way. Which genealogical and areal groupings are chosen as the sampling groups in the stratification naturally varies from one study to another. Different genealogical classifications and areal divisions can be chosen as the basis, and they can be used at different levels of classification – large families vs. branches lower in the trees, or large continent-size areas vs. smaller linguistic areas. In the DV method stratification can be done on the basis of any classification that is representable in tree format (dendrogram); most often, genealogical trees are used. In the GM method, stratification follows primarily the groupings known as genera and macroareas initially proposed by Dryer (1989); these will now be briefly discussed.

The GENUS, as defined by Dryer (1989, 1992, 2005a, 2013), is a level of genealogical classification intended to be comparable across the world in terms of time depth. The time depth of genera is not more than 3,500 to 4,000 years (Dryer 2005a: 584). Earlier, Bell (1978: 146–149) had introduced language groups with a time depth of not more than 3,500 years as a basis of stratification. Familiar examples of genera are the branches of Indo-European: Germanic, Romance, Slavic, etc. The time depth of, e.g., Germanic is much less than 3,500 to 4,000 years since Germanic languages split from proto-Germanic much more recently (see, e.g., Henriksen & van der Auwera 1994: 1). But there are no other languages with which the Germanic languages have a common ancestor dating

back to less than 3,500 to 4,000 years. The split from the other Indo-European branches is older than that and therefore these belong to genera different from Germanic. In many areas of the world, genera are the maximal level of grouping whose genealogical relationship is uncontroversial. Genealogical classifications are far from being agreed on in many parts of the world, and many higher level classifications are rather uncertain, sometimes based on geography more than genealogy. This is often due to insufficient research on languages in areas such as Australia, New Guinea, and South America (even though a lot of progress in genealogical classification has recently been made in these areas). Therefore, the family trees proposed for languages in different geographical areas are not directly comparable and remain somewhat incommensurable for the purposes of typological sampling. According to Dryer's (1989) proposal, as these problems mostly involve levels higher than the genus, the choice of the genus level as a basis for stratification should reduce the effects of the problems of classification (the choice of the genus level as the basis of stratification is addressed in more detail in Section 4, in which the GM method is described). On the negative side, as also admitted by Dryer (2000: 348–349), the criteria of what counts as a genus cannot be applied mechanically to produce an objective list of genera. Dryer's classification is based on existing literature and on views of experts in different families and geographical areas. These are, however, the sources that any worldwide classification such as the one found in *Glottolog* (Hammarström et al. 2015, henceforth GLOT) has to rely on, and in this sense Dryer's list of genera is methodologically comparable with other worldwide classifications.

Dryer's (2013)[5] classification in *WALS* contains 521 genera in total (in addition to these, the classification contains two groups of languages that are not considered genera in the genealogical sense: creoles and pidgins, and sign languages). More generally, the structure of Dryer's genealogical classification is as follows. Each language belongs to a genus and each genus belongs to a family (a language can be the only member of its genus, and a genus may form a family on its own). A family is defined as "the highest level of classification widely accepted by specialists" (Dryer 2005a: 584). The number of families in the classification is 215. In some cases, there is an intermediate level of classification – subfamily – between the family and genus levels; however, these are not taken into account systematically, but provided only in cases the subfamily is well known. The classification thus has three to four levels: family – (subfamily) – genus – language. As a result, the tree of every family has the same rather flat

5 The current version of the classification available at http://wals.info/ contains some minor corrections with respect to the 2013 version thta we have used as basis for our simulations in Section 5 and that we refer to here.

structure, and the classification is not meant to represent everything that is known about the genealogical relationships between languages within families.

MACROAREAS, according to Dryer (1989), are continent-size linguistic areas which are independent of each other, but within which languages are to some extent typologically similar due to either (ancient) contact or (very deep) genealogical affinity, beyond the reach of the methods of historical linguistics. The following six macroareas are distinguished by Dryer (1992): Africa, Eurasia, Southeast Asia & Oceania, Australia & New Guinea, North America, and South America. The boundaries of macroareas mostly follow geographical divisions, but sometimes deviate from these. Families are not usually split between two macroareas, even where they contain languages belonging to two different geographically defined continents. Thus, all Semitic languages, even the ones spoken in Asia, belong to the African macroarea with the rest of Afro-Asiatic, and the Chibchan languages of Central America belong to South America where most of the Chibchan languages are situated (cf. Dryer 1989: 268). In these cases, the relevant part of the continent is occupied exclusively by languages belonging to the genus/family in question. The Austro-Asiatic family, on the contrary, is split between two macroareas: the Munda genus is geographically part of Eurasia surrounded by other Eurasian genera and typologically similar to these, whereas the rest of the family is geographically part of Southeast Asia & Oceania, and typologically similar to the languages of this macroarea.

Obviously, the mutual independence of the macroareas is not an unproblematic issue, and some genealogical relationships and contact influences may surely be found across the boundaries of macroareas. Dryer (1989: 268) takes their effect to be sufficiently small that it can be ignored in the method of testing universals he proposes. For the present purposes, as the requirement of the independence of sampling units is not as crucial for variety sampling as it is for probability sampling, the possible dependencies across the borders of macroareas would be less significant. It may also be noted that macroareas are traditional linguistic areas, albeit very large, in the sense that they are (partly) defined by shared linguistic features. There is thus a potential danger of circularity if one first defines an area on the basis of typological similarity and then uses it as a basis of stratification in typological sampling. Hammarström & Donohue (2014) review Dryer's six macroareas and propose a somewhat different division based entirely on geographical independence without reference to linguistic data: Africa, Eurasia, Multinesia, Australia, North America, and South America. These areas (renaming Multinesia as Papunesia following the *Glottolog*) have been adopted in the latest editions of *WALS* instead of Dryer's original six areas. Bickel & Nichols (2013) provide a worldwide areal classification with two levels: 10 continent-size areas and 24 smaller-scale ones. Their

classification is based on assumptions about contacts in history, informed by historical, genetic, anthropological, and archaeological knowledge, but to avoid circularity in areal linguistic research, linguistic information is not used in defining the areas. The continent-size areas could be used as an alternative to Dryer's macroareas in areal statification. They are: Africa, West and Southwest Eurasia, North-Central Asia, South and Southeast Asia, New Guinea and Oceania, Australia, Western North America, Eastern North America, Central America, and South America. Since these areas are defined without using linguistic information, it is an empirical matter to prove or disprove their validity as linguistic areas. So far, no information is available as to how well these proposed areas actually capture areal-typological patterns, but at least the following problem may be pointed out: the classification divides the Circum-Baltic linguistic area (see Koptjevskaja-Tamm & Wälchli 2001) between two separate continent-size areas: West and Southwest Eurasia vs. North-Central Asia. In this article, we have used Dryer's original macroareas in discussing the GM method in Section 4 and in simulating it in Section 5, following the original idea of macroareas as continent-size linguistic areas.

2.2 Sampling methods in the literature

This section provides an overview of sampling methods proposed in prior literature. The focus will be on methods that are presented as explicit, systematic, and clearly delimited procedures applicable as such and whose application results in a language sample. None of them is explicitly designed as a method of variety sampling, but their potential usability in variety sampling will be addressed in the discussion.

Some authors, e.g, Bell (1978) and Maslova (2000), have discussed sampling issues at length, but have not proposed a method in the above sense, and although their work comes up at various points in this article, they will not be discussed as actual sampling methods here. Perkins (1989) proposes a method of calculating ideal sample size and representation of each genealogical grouping. However, as pointed out above, this method is dependent on the crosslinguistic variables investigated and can be applied only when the variables are known beforehand. It is therefore not optimal for the present purposes. Perkins (1992), paying attention to the relationship between language and culture, stratifies his sample primarily on the basis of cultural groupings, but less important roles are played by genealogical and especially areal groupings – the most important grounds for stratification in the present context. Dahl (2008) measures typological distances between languages on the basis of the *WALS* database, and uses

these measures to build a sample in which only languages that are sufficiently divergent typologically are included. The areal breakdown of the languages in the resulting 101-language sample is as follows: Africa 16, Europe 8, Asia 17, Oceania 3, New Guinea 9, Australia 9, North America 19, and South America 20. The number of languages sampled from each area is then taken as a measure of the internal diversity of the area. Dahl offers the general advice that in typological samples continents should be represented in approximately the proportions suggested by his diversity measures. He does not, however, offer a codified method of constructing a variety sample that could be tested here.

Tomlin (1986) collected data on word order in 1,063 languages without a systematic method for selecting the languages. From these languages he then took a subsample of 402 languages, seeking to represent each genealogical and areal grouping recognized in his frame in proportion to the number of languages the grouping contains. The principles of post-hoc sampling used by Tomlin could also be used for *a priori* sampling and it is therefore relevant to discuss them here. In Tomlin's sample, small genealogical groupings, on the one hand, are grossly underrepresented, being lumped together to create larger groupings for the purpose of sampling, which results in leaving out a large number of small groupings altogether. Large groupings, on the other hand, are clearly overrepresented with respect to what would be optimal in a sample seeking to maximize diversity. The sampling method may capture the distribution of attested languages quite well, but it is not well-suited for variety sampling.

Dryer's (1989, 1992) method of sampling is similar to Tomlin's in the sense that languages are first included in the sample without a systematic method of selection, and this bottom-up approach is then complemented by a more systematic stratification at the stage of testing generalizations. The number of languages in the database is well over a thousand (cf. the number of languages in Dryer 2005b), but not all languages are coded for all features. The size of the total sample thus differs from feature to feature. When testing generalizations, Dryer lumps together the languages of each genus and counts genera instead of languages. A pattern is taken to be universally valid if it is preferred in the majority of genera in all macroareas. This may be a useful method of testing crosslinguistic generalizations, but it does not offer a sampling procedure of the type aimed at here. The same is to a large extent true of the post-hoc sampling method for genealogical control proposed by Bickel (2008).

The sampling method used by Nichols (1992) can also be characterized as a bottom-up approach, where sample size and the number of languages coming from different groupings are not determined beforehand. The world is divided into ten sample areas whose boundaries do not touch each other. The representation of the genealogical diversity of each sample area is to be guaranteed

by including a language from each independent family and from each subfamily of large families with significant time depth (the number of subfamilies is limited to six for each family, thus only six subfamilies of Indo-European, for example, are present in the sample). The sample includes 174 languages. Nichols's sampling method is designed for the purpose of revealing worldwide areal distributions of linguistic features and is clearly not a general method of variety sampling, one specific problem being that the geographic areas falling outside the boundaries of the sample areas are not represented in the sample at all.

Bybee et al. (1994) adopt a top-down sampling procedure where the representation of each genealogical grouping of the world's languages is determined by its size and internal structure. No areal stratification is used. In Voegelin & Voegelin's (1977) classification, which they use, there are 55 minimal groups, i.e., isolates or families with fewer than 21 languages, and 24 other families (with more than 20 languages each). Aiming at a sample of 75 to 100 languages (cf. Perkins's recommendation above), they choose the following principles to calculate the number of languages to be taken from each grouping. Only two languages are selected from the entire set of minimal groups. The larger families are all represented by at least one language: one language is selected from each family with up to 40 members, and the representation of each family with more than 40 languages is determined by the structure of the family tree (see Bybee et al. 1994: 303–310 for the technical details). The theoretical sample size is 94 languages, but in cases adequate sources are unavailable, languages are left out, and the actual sample size is reduced to 76. Our main objections to the use of this sampling method as a general method for constructing variety samples are the following. Firstly, just like in Tomlin's sample, isolates and small families are underrepresented with respect to the potential variety they bring into the sample – from the point of view of variety sampling, the size of a grouping does not constitute grounds for regarding it as less important in the overall picture of the world's linguistic diversity (cf. the discussion on the effect of extralinguistic historical accidents on the actual language population). Secondly, the numbers of languages to be taken from each grouping are chosen in view of a sample size of 75 to 100 languages, and it remains unclear how the method should be adjusted for other sample sizes. Thirdly, there is no areal stratification.

Although the sampling methods taken up in this section have not been designed explicitly for variety sampling, we addressed their potential strengths and weaknesses in this function. Our conclusion will be that none of the methods show enough potential as generally applicable methods of variety sampling, and they will therefore not receive further attention in this article.

We will now move on to discussing two methods that have been especially designed for variety sampling: the DV method and the GM method.

3 The Diversity Value sampling method

The Diversity Value (DV) method proposed by Rijkhoff et al. (1993) and Rijkhoff & Bakker (1998) is explicitly designed as a method of constructing variety samples, and it is therefore chosen as one of the methods to be tested in Section 5. This section will provide a short introduction to the method; the introduction is kept short because the method has been extensively discussed elsewhere (cf. Bakker 2011).

The DV method takes genealogy as its basis of stratification. Any classification of the world's languages that can be represented in the form of a family tree can be used. The diversity value (DV) of a genealogical grouping (independent family or lower-level grouping) is a measure of its internal diversity. To put it simply, the DV of each genealogical grouping is computed on the basis of the structure of the family tree (number of non-terminal nodes in the tree). More specifically, for each node in a tree, and starting with the top node, its DV is determined on the basis of the complexity of the tree under it, recursively calculated in terms of the number of levels under it and the number of sister nodes on each lower level. To higher levels in the tree greater weights are assigned than to lower levels. The number of terminal nodes, i.e., languages, is disregarded. If the desired sample size is equal to the number of independent families in the classification, one language is taken from each family. If the sample size is smaller than the number of families, the DVs of the families determine the probability each family has for being represented in the sample. If the sample size exceeds the number of families, then the number of languages selected from each family is determined on the basis of the DVs of the respective families. Per family, these languages are then recursively assigned top-down to the lower levels of the family tree, proportionally to their DVs. This process stops when a number of sister nodes has been reached that is smaller than or equal to the number of languages assigned to their mother node. Determining the number of languages for each genealogical grouping on the basis of its DV aims to guarantee that the internal variety of each grouping is represented in the sample as well as possible. It is assumed that the complexity of the tree rather than the number of languages, is a good indicator for potential variety. It is further assumed that higher ("older") splits in a tree contribute more to the variety among the languages under them than lower ("younger") splits.

In the remainder of this section, we will take up some potential problems of the DV method, precisely since it has been used quite frequently so far, and the GM method discussed further on has been developed partially to try and counter some of these potential problems. Firstly, it may be noted that the method itself involves no areal stratification. It is true that since the method can be applied to any classification that can be represented in the form of a tree, it could be applied to an areal tree as well. This would, however, mean replacing genealogical stratification by areal stratification rather than combining them, and in any case no worldwide areal classification in the form of a tree is available.[6] Bakker (2011: 118) mentions that with the computer programme generating DV samples, a genealogically based sample may be areally stratified to the extent that areal information is available for the languages in the classification. This would be a more feasible approach to taking areality into account in DV sampling.

Problems may also be caused by the fact that counting DVs relies on the details of the genealogical classification chosen, and the whole method is thus affected by the uncertainties and inconsistencies in genealogical classifications (cf. the discussion above). A concrete example of these problems can be seen in Rijkhoff & Bakker (1998: 289–291), where they compare the samples created by their method on the basis of two different classifications – Ruhlen (1991) and the 13th edition of the *Ethnologue* (Grimes (ed.) (1996), Grimes & Grimes (1996); henceforth E13). Although the samples based on the different classifications have more or less equal numbers of representatives for many families, certain families (or proposed super-families) are represented by very different numbers of languages in the samples based on the different classifications. To take the difference between their alternative 200-language samples as an example, the numbers are as follows: Afro-Asiatic: 13 languages in the sample based on Ruhlen and 8 languages in the sample based on E13, "Amerind": 42 (Ruhlen) vs. 67 (E13), Australian: 16 (Ruhlen) vs. 9 (E13), Sino-Tibetan: 9 (Ruhlen) vs. 6 (E13). Contrary to what Croft (2003: 21) thinks, these differences are not insignificant.[7]

6 Bickel & Nichols's (2013) classification into 10 continent-size and 24 smaller-scale areas can of course be represented as a tree, but the tree will be very flat with only three levels (larger area, smaller area, language) and applying the DV algorithm to such a flat classification would not bring much added value.

7 It should also be noted that the criticism of under- and overrepresentation that Rijkhoff & Bakker (1998: 299–303) express against Dryer's (1992) and Stassen's (1997) samples is somewhat unfounded. The latter two are examples of samples that have actually been used in typological studies, and they are thus affected by the poorer availability of sources in some areas. Therefore, the ideal sample computed by Rijkhoff & Bakker is not directly comparable to theirs.

The problems of classification can also be seen within a given worldwide classification. It is easy to observe incommensurabilities between trees of different genealogical groupings in one and the same classification. These may be due, e.g., to insufficient research on genealogical relationships in many parts of the world. But even in cases of relatively well-studied families, the trees of two genealogical groupings may turn out to be incommensurable: two groupings may have fairly equal numbers of nodes but these nodes represent a much greater time depth (and thus potentially greater linguistic diversity) in the case of one of the two groupings. To take an example from the classification of the *Ethnologue* (18th edition, Lewis et al. (eds.) 2015 henceforth E18), the genealogical grouping "German" (a branch of High German [< West Germanic < Germanic < Indo-European]) has four levels of classification below it and is (in this vertical sense) equivalent to the entire Uralic family which also has four levels of classification below the top node, both thus having three levels if the terminal nodes are disregarded. The difference in time-depth between German and Uralic is several thousand years. It is true that the widths of the two groupings are different, and a low-level subgroup would not be directly compared to an independent family in the DV method (lower nodes contribute less to the overall DV than higher ones), so the DVs of German and Uralic would not be exactly equal. However, note also that the German node still has four nodes above it, and the whole Indo-European tree is thus much more diverse than its Uralic counterpart even though Uralic is generally thought to be at least as old as Indo-European – the most common estimates of the age of Proto-Uralic range between 5,000 and 7,000 years before present (see Janhunen 2009: 65–68). Another example can be taken from Siouan-Catawban, which is an independent family in the *Ethnologue* classification: it has six levels of classification below the top node and is thus two levels deeper than Uralic which has only four. The age of Proto-Siouan-Catawban has been estimated at 4,000 years (Rankin 1993, cited in Campbell 1997: 142). These examples show how DVs counted on the basis of nodes in family trees can lead to wrong interpretations of the diversity of families (see also Croft 2003: 21–22). To deal with these kinds of problems, the DV method can in principle weigh the nodes according to their time depth (see Bakker 2011: 118). However, information on time depths cannot be systematically added to all nodes in all families, since it is not available in the classifications on the basis of which the samples can be generated – to use cladistic terms, the trees are simple cladograms, not chronograms in which the branch spans would indicate time depths.

Despite these potential problems – the lack of areal stratification and the assumed unreliability of genealogical trees, especially when used as a stand-alone basis for sampling – the DV method has a number of obvious advantages.

It takes full advantage of the diversity expressed in genealogical and other types of trees, it is fully explicit and formalized, and has been implemented in the form of a computer programme. As a result, it provides reproducible samples comparable across different studies. Indeed, over the past two decades, many typological studies have relied on this method for establishing their samples.

4 The Genus-Macroarea sampling method

This section discusses the Genus-Macroarea (GM) method proposed by Miestamo (2003, 2005). The treatment will be more extensive than that of the DV method in Section 3, since the GM method has not been as prominently present in the literature on sampling, and we will also introduce some new features to the method. In the GM method, the primary genealogical stratification is made at the genus level, and the primary areal stratification at the level of macroareas. The following subsections will show in more detail how this is done. Section 4.1 deals with a bottom-up variant of the method in which sample size is not predetermined. In Section 4.2, a top-down variant in which sample size is determined in advance will be discussed.

4.1 GM sampling without predetermined sample size

In the GM method, constructing a variety sample without predetermined sample size will, at its simplest, mean picking one language from every genus. In principle then, every genus is represented in the sample. As different geographical areas will be unevenly represented due to the poorer availability of data in some areas, an additional component to the method, to be introduced further below, will make it possible to achieve a better balance between macroareas. The method thus involves different samples or levels of sampling: the Genus Sample, the Core Sample, the Restricted Sample, and the Extended Sample. These will be introduced in turn below.

To achieve maximal representation of the world's linguistic diversity, the method tries to include languages from as many genera as possible. The aim is thus to sample one language from every genus. The choice of a language from the list of languages in each genus should preferably be made randomly. Since random selection is not crucial for the goals of variety sampling, one may alternatively choose the languages on the basis of the availability of the sources. The sample containing one language from every genus will be called the GENUS

SAMPLE (GS). Using the *WALS* list of genera (Dryer 2013), this will result in a GS of 521 languages.[8]

The use of the genus level has certain advantages. The very idea behind the notion of genus is that these genealogical groupings should be crosslinguistically comparable in terms of time depth, which makes them very useful in typological sampling. The time depth of genera is such that languages belonging to different genera should be sufficiently far removed from each other to maximize the potential variety in the sample while still enabling the rather large sample size required in variety sampling. It is true that closely related languages or even dialects of one language can show typological differences with respect to a given linguistic feature, but it remains the case that typological differences are far more probable between unrelated or less closely related languages. Thus, a good variety sample should have a representative from as many of these relatively independent groups as possible.

Dryer (1989) proposed the notions of genus and macroarea to be able to test universal generalizations such as the Greenbergian word order correlations. In other words, these concepts were originally designed for the purposes of probability sampling. Dryer used them to stratify his sample, to maximize the representativeness of the sample and the independence of the sampling units. As discussed in Section 2.1, representativeness and independence are important goals in variety sampling as well, although in somewhat different ways, and genera and macroareas provide a feasible and promising basis of stratification for variety samples, too.

Sometimes a typologist using a language sample encounters the question whether the language picked from a given genealogical grouping (genus in this case) is a good representative of that grouping, whether, e.g., English would be a good representative of Germanic. Such questions are based on a misunderstanding of what sampling is about. Typologists working with samples aim to draw crosslinguistic generalizations and do not make claims about individual

8 The *WALS* list of genera is not a full list of genus-level groupings in the world. It only includes those genera for which some languages are present in the *WALS* database. Especially for South America and New Guinea, the *WALS* list is not complete, as becomes clear by comparing the *WALS* classification and the GLOT classification in these areas. As the selection of languages present in *WALS* is heavily influenced by the availability of descriptive materials (only those languages for which some data exists can be included in the research behind the maps), this genus list covers quite well the genera for which described languages can be found. One is naturally free to add genera to the list, e.g., if data is found on a language of a genus not present in *WALS*. Like any classification, the list of genera is not the final word about the genealogical relationships between the world's languages; it will undergo improvements as research progresses.

genera. If one wants to say something about Germanic, then one studies several Germanic languages, and does not use general typological samples. Whether a given language is a good representative of its genus is less important for the purposes of large-scale sampling. In a large sample such effects will not affect the overall crosslinguistic distributions to a significant degree. If one wants to take genus-internal variation into account, and to select more than one language from a genus, one will have to accept a significant increase in the size of the sample, but as discussed above, extra-large samples are not feasible in most research settings. However, there is a sense in which the typicality of a language of its genus may be relevant in the context of variety sampling: a language may be atypical of its genus because it has acquired features from neighbouring languages belonging to other genera, and it may therefore be less different from its unrelated neighbours than are its genus mates.[9] If one is not picking languages at random, one may take this issue into account to the extent that information about typicality and contacts is available.

As noted above, the genus level, as originally proposed by Dryer (1989), is meant to be a fairly uncontroversial level of classification and comparable across the world. On the assumption that the genus level has these properties, the use of genera as a stratification basis could help us to overcome the problems caused by the uncertainties and incommensurabilities in classifications addressed in Section 3. As discussed above, the notion of genus has its own problems, and has been critized for there not being fully objective criteria for deciding what counts as a genus; this creates a potential problem for the GM method. Note, however, that the sampling principles proposed here can be used with other classifications following a similar goal of worldwide comparability, if such classifications become available in the future. It is the worldwide comparability of groupings that is crucial for choosing the classification used with the method.[10] The simulations of the GM method in Section 5 put the usability of the genus classification for sampling into test.

9 We thank an anonymous referee for pointing this out.

10 A referee pointed out that the 328 stocks identified by Bickel & Nichols (2013) could be used as an alternative to genera in this kind of sampling. The stock is meant as the highest level of classification that can be demonstrated and reconstructed. Their advantage would be that they are based on more objective criteria than genera. However, they are not comparable in terms of time depth. In areas where less work has been done and where the comparative method has not been as successful as in the case of Indo-European, the proposed stocks tend to have a much lower time depth than in areas where comparative work has been more successful in finding deeper relationships. This naturally applies to the highest levels of classification in E18 and GLOT as well.

In addition to the 521 genera, Dryer's list treats pidgins and creoles as well as sign languages as independent groups on a par with genera. The choice of how to treat these non-genus groups in GM sampling is somewhat arbitrary, but as a default, one creole, and, if there is no explicit focus on spoken languages, one sign language can also be included in the sample at this stage. Pidgins, however, are not the native language of anyone and do not count as natural languages in this sense, so they can be left out of the sample. This will raise the number of languages in the GS to 523.

In an ideal world, one could find data for all languages in the GS and include them all in the database one is compiling, but this is difficult in practice: for many of the languages, adequate sources may not exist or they may not be available in libraries or online – or if available, they may not discuss the phenomenon under study in sufficient detail, if at all. When adequate sources are not found for a randomly selected language, one may choose another language from the same genus, again randomly or by picking the closest relative of the previously chosen language if genealogical information is available, and see if adequate sources can be found for that language. If needed, this process may be repeated until one finds a language with sources usable for finding the relevant data. If no languages with sufficient sources can be found for a genus, then that genus cannot be included in the study. The set of languages (one per genus) that one is able to include in the study in this way will be called the CORE SAMPLE (CS). Depending on the research question and the sources available, the size of the CS may be much smaller than that of the GS. To take an example from the study for which this sampling method was used for the first time, out of a total of (then) 412 genera, Miestamo (2003, 2005) was able to include a language from 239 different genera plus one creole – for the rest, sources were not available, or they were not suitable for answering the research questions. The CS thus had just 240 languages.[11] The size of the CS depends very much on the research question. Topics that require one to delve deeper into the grammar of each language and that therefore require thick and thorough grammars or specialized studies on the topic to be used as sources do not allow CSs as large as topics for which sources are easier to find. In any case, even in studies in which extensive CSs cannot be built, it is desirable to try to cover as many genera as possible.

Using this method will always result in a CS, but the CS may optionally be extended or restricted for different purposes. The restrictions will be discussed first. Since the number of genera and languages included in the sample from

11 The *WALS* classification was not yet available at that time and the list of genera used was based on an earlier version of Dryer's classification.

each geographical area depends on the availability and quality of the sources and since some areas are better studied than others, some areas will be better represented in the CS than others. The bibliographic bias thus tends to introduce an areal bias to the CS.[12] This is illustrated by the numbers in Table 1. The table shows, for each macroarea, the numbers of genera and the numbers of languages included in the CS in Miestamo's (2003, 2005) study,[13] as well as the percentage of the number of genera covered by the included languages ("RS" in the rightmost column stands for the Restricted Sample to be discussed below). Better-described areas are overrepresented in the CS (most notably Eurasia with a coverage of 97% of its genera) with respect to the less well described areas, especially Australia & New Guinea (with a coverage of only 43.2%) and South America (only 44%). The percentages of coverage may look rather low for these areas, but the bibliographic bias affects any sample of this size in the same way and the sample does not compare unfavourably with other typological samples in this respect. The availability of sources has seen significant improvement during the last few years, so that repeating the same study today would give a much better coverage and a larger CS. Hammarström (2009) provides estimates of the bibliographic coverage of different families and macroareas.

Table 1: Genera and languages in CS and RS by macroarea; adapted from Miestamo (2005: 36).

	Genera	Languages in CS	Coverage	Languages in RS
Africa	66	45	68.2%	29
Eurasia	35	34	97.1%	15
Southeast Asia & Oceania	49	26	53.1%	21
Australia & New Guinea	88	38	43.2%	38
North America	83	56	67.5%	36
South America	91	40	44.0%	39
Total	412	239	58.0%	178

The CS is first and foremost a variety sample. The areal bias is not directly harmful for the general aims of variety sampling, but if one wants to make some quantitative generalizations as well, e.g., get a better idea of global frequencies of linguistic features, one should try to remove this bias to the extent possible. A method of achieving a better areal balance was introduced in Miestamo

12 One may of course argue that there are genealogical biases in the CS as well if one emphasizes the significance of the family-level: of two families comprised of several genera, one may get all or most of its genera represented while the other may get only one or two.

13 Remember that this study used a pre-*WALS* list of genera.

(2005). The RESTRICTED SAMPLE (RS) is a subsample drawn from the CS with the aim to balance the representation of each macroarea. In the RS, each macroarea is represented in proportion to its genealogical diversity, i.e., to the total number of genera in the macroarea. This even distribution is reached by randomly eliminating languages/genera from each of the better represented areas until the percentage of the number of languages/genera included from each area is equal to the least well represented area. In our example case, the least well represented area is Australia & New Guinea with 43.2% of its genera covered in the CS. The RS will thus include 43.2% of the total number of genera in each macroarea: 43.2% of the total number of genera in Africa (66) gives the number of African languages in the RS as 29, 43.2% of the total number of genera in Eurasia (35) gives the number of Eurasian languages in the RS as 15, etc. for all six macroareas (see the right-hand column in Table 1). 43.2% of the world total of genera (412) gives the size of the RS as 178 languages. The coverage of the least well-represented area thus defines the maximal size of a RS that can be drawn from a given CS; a lower percentage may also be chosen, if a smaller sample is needed for some reason.

The non-genus groups, creoles and sign languages, which cannot be assigned to a specific macroarea, are not part of these calculations. As noted above, the choice of how to treat creoles and sign languages is somewhat arbitrary, but as a default one creole and one sign language may automatically be included in the RS (in Miestamo (2005), one creole was included and the actual sizes of the samples were 240 for the CS and 179 for the RS).

The RS avoids the areal bias found in the CS by rendering the representation of each macroarea proportional to its internal genealogical diversity, i.e., for each macroarea, the same percentage of the total number of genera gets represented. With the areal bias removed, the RS is better suited to serve as a basis for quantitative analysis. It should be stressed that this methodology is primarily intended for variety sampling, but restricting the sample this way makes the quantitative treatment of the data more reliable, if one wants to include a quantitative component in a primarily qualitative study.

Just as other genealogical groupings than Dryer's genera could be used if alternatives meeting the requirement of comparability were available, other macroareal classifications could also be chosen for areal stratification. As pointed out in Section 2.1, the continent-size areas proposed by Bickel & Nichols (2013) or the six macroareas proposed by Hammarström & Donohue (2014) could be used as such an alternative.

The method of restricting the sample used in Miestamo (2005) did not include any means to ensure that the languages that are left outside the RS come from different families. To take an example from Eurasia, the RS contains

four Indo-European languages but some Eurasian isolates and small families are left out, e.g., Ainu and Yukaghir. Possible ways to solve this problem will now be discussed. The numbers of languages that have to be left out from each macroarea (= n_m) are counted from the difference between percentages of coverage in the CS and the RS. A rather simple option would be to take the n_m best-represented families in each macroarea and choose (randomly) one language from each of these families to be left out of the RS. If n_m is bigger than the number of families with more than one language in the sample, then one should start a second round from the best-represented family. Another option is to recursively select the family (or families) with the greatest number of languages in the sample in each macroarea, and leave out a language from that family (or those families), until n_m languages have been removed. The problem with the former solution would be that languages would be taken out from small families as easily as from bigger ones and their proportional representation would then decrease much quicker than that of bigger families. The latter solution is therefore to be preferred.[14]

As to areal balancing, a similar method can be used to make sure that the languages to be left out come evenly from different subareas of each macroarea. For this, an adequate worldwide classification of linguistic areas below the macroareal level is needed. Despite the doubts expressed by Stassen (1997: 7) and Wichman & Kamholz (2008: 251–252), a few possible candidates are available. Murdock's (1968) areas are primarily designed for sampling in cultural anthropology and could in principle be used as a basis of stratification against cultural biases, but they do not translate into linguistic areas as such; furthermore, being as many as 200, they provide a more fine-grained grid than would be usable for the present purpose. Tomlin's (1986) 26 sampling areas are meant for language sampling; they are established by first taking the existing non-controversial linguistic areas proposed in the literature, and then defining the remaining areas either negatively with respect to the established areas or by paying attention to geographical or political boundaries, and ultimately to latitudes and longitudes. The most up-to-date alternative to be used as a smaller-scale areal classification in worldwide sampling is provided by the 24 smaller-scale areas proposed by Bickel & Nichols (2013), see Section 2.1.

14 A similar procedure could be applied within families from which more than one language has to be suppressed, to make sure that the representation of intermediate levels of classification such as major branches of the families remains balanced and no major branch is deleted from the sample if another branch at the same level still has more than one representative. However, since the intermediate level, subfamilies, is not systematically given in the *WALS* classification, this can be done with some families only.

Alternatively, the finer geographical balancing could be achieved by paying attention to geographical distances without any division into subcontinental areas, and building in a component that would ensure that the languages chosen within a macroarea are sufficiently far removed from each other.

The construction of a RS according to this method will now be illustrated in detail using an existing dataset, namely the word order data in Dryer (2005b); the illustration does not involve areal stratification below the macroareal level. The dataset includes 1228 languages from 378 genera (here the division into genera is the one used in the 2005/2008 edition of *WALS* with a total of 475 genera; Dryer 2005).[15] Suppose that these data had originally been collected using the GM sampling method. The size of the CS would then be 378 languages (and the total of 1228 languages would make an Extended Sample (ES) which basically means the CS plus all additional languages included in the study, see below). The macroarea with the lowest representation of its genera in the dataset is South America, with 59% of its genera represented. The largest RS that can be drawn from the dataset thus contains 280 languages (59% of the 475 genera). Selecting a RS of 280 languages means including 59% of the genera in each macroarea: Africa 41 (out of 70), Eurasia 23 (39), Southeast Asia & Oceania 27 (45), Australia & New Guinea 79 (134), North America 54 (92), South America 56 (95). Of the 70 African genera, 67 are represented by at least one language in the *WALS* 81 dataset. To arrive at the 41 languages to be included in the RS, 26 African genera thus need to be deleted. They are taken from the best-represented families: Nilo-Saharan with 25 (out of its total of 25) genera represented, Niger-Congo with 24 (out of its total of 25) genera represented, and Afro-Asiatic with 13 (out of its total of 13) genera represented. Deleting 13 Nilo-Saharan genera, 12 Niger-Congo genera, and one genus from Afro-Asiatic will eliminate the required 26 genera from Africa and leave 12 genera from each of these families. This is expressed more schematically for all macroareas as follows:

(i) for Africa 67(70)–26 = 41: Nilo-Saharan 25(25)–13 = 12, Niger-Congo 24(25) –12 = 12, Afro-Asiatic 13(13)–1 = 12;
(ii) for Eurasia 38(39)–15 = 23: Indo-European 10(10)–8 = 2, Nakh-Daghestanian 4(4)–2 = 2, Dravidian 4(4)–2 = 2, Uralic 3(3)–1 = 2, Altaic: 3(3)–1 = 2, Chukotko-Kamchatkan 2(2)–1 = 1;
(iii) for Southeast Asia & Oceania 39(45)–12 = 27: Austronesian 16(17)–8 = 8, Sino-Tibetan 11(14)–3 = 8, Austro-Asiatic 8(9)–1 = 7;

15 The number of genera per macroarea in the 2005/2008 edition is as follows: Africa 70, Eurasia 39, Southeast Asia & Oceania 45, Australia & New Guinea 134, North America 92, South America 95.

(iv) for Australia & New Guinea 105(134)–26 = 79: Australian 29(34)–18 = 11, Trans-New Guinea 19(22)–8 = 11;
(v) for North America 73(92)–19 = 54: Penutian 9(10)–6 = 3, Uto-Aztecan 8(10) –5 = 3, Oto-Manguean 7(9)–4 = 3, Hokan 5(8)–2 = 3, Salishan 3(5)–1 = 2, Algic 3(3)–1 = 2;
(vi) and for South America 56(95)–0 = 56.

Note that in cases where the required number of eliminated genera leaves one or more of the affected families with fewer genera than the others, the families with a lower total number of genera are affected first. Thus in Eurasia, one of the affected families is left with only one genus, and this is the smallest one, namely Chukotko-Kamtchatkan. This has a minimal effect, but if a choice has to be made, it can be made according to the diversity of the families rather than just randomly. Once the numbers of genera to be deleted from each macroarea and family have been determined, one proceeds to choose (randomly) which ones of the genera are eliminated and which ones stay in the RS.[16]

In addition to drawing restricted subsamples from the one-language-per-genus CS, the CS can also be extended by including one or more additional languages from one or more genera. These additional languages together with the languages included in the CS form the EXTENDED SAMPLE (ES). The CS is thus a subset of the ES. There are various reasons why one might want to include more than one language from some genera. Although the CS should usually be enough to bring out the general crosslinguistic variation in the domain of inquiry, extending the sample beyond the CS has the effect of increasing the coverage of the sample, and it thus has a positive effect on the linguistic variety captured by the sample. Additional languages may be included simply because one may already have looked at the domain under study in them or may have easy access to the data.[17] If one encounters rare types or other interesting phenomena in a language of the CS, taking more languages from the same genus may increase linguistic variety in the sample. Looking at closely related languages may also increase variety in the sense that variants of one and the same type or types intermediate between different types may then be better captured (cf. Croft 2003: 22). In the case of very large

16 Since this dataset was not originally compiled with the present method, we do not know which language in each genus would belong to the CS and will eventually make it to the RS. We still have to choose, again randomly, the language to be included in the sample from each genus that is represented by more than one language in the dataset.

17 In Miestamo (2003, 2005), the ES included 57 languages in addition to the CS and the size of the ES was thus 297 languages. The reason for including these additional 57 languages was that they had been analysed for the purposes of a related project, namely *WALS*, and they could therefore be easily included in the study adding thus to the diversity of the overall sample.

genera (e.g., Bantoid, Oceanic, Pama-Nyungan) that cover a large geographical area, aiming at better areal coverage by including several languages from these large genera may increase variety.[18] Naturally, in terms of (genealogical) diversity, an ES of a given size in which some genera are represented by more than one language compares negatively to a hypothetical CS of the same size where each language comes from a different genus.

So far, we have seen that the samples produced by the GM method are quite extensive. However, smaller samples may be needed for various purposes, e.g., for pilot studies, and indeed, the method may be criticized for being unable to produce them. To satisfy this need, a variant of the method allowing for any predetermined sample size, following the same principles of areal and genealogical representation, has been developed, and is introduced in the next subsection.

4.2 GM sampling with predetermined sample size (the top-down variant)

This section will discuss a top-down variant of the GM sampling method, in which the size of the sample is determined in advance. Any sample produced with this variant of the method will be called a PRIMARY SAMPLE (PS). The proportional representation of the genealogical diversity of each macroarea is counted using the percentages that the number of genera in each macroarea represents of the total number of genera in the world, and the number of languages to be included in the PS from each macroarea is then counted from the predetermined sample size by using these percentages. As in the RS above, genealogically more diverse areas are represented by a higher number of languages in the PS than areas that show less diversity. This avoids the overrepresentation of, e.g., the Eurasian languages and the underrepresentation of areas such as Australia & New Guinea and South America, typical of so many language samples (cf. also Dahl 2008). A 50-language PS produced by this method was used in Miestamo (2009).

18 The reasons for extending the sample discussed so far have to do with increasing variety, but building an ES may be motivated by other goals as well. For example, to address diachronic and/or contact issues, and to examine areal patterns, a closer look at specific areas or genealogical groups may be needed to supplement the general typological survey (cf. Stolz & Gugeler 2000: 55–60).

This principle will now be illustrated for different sample sizes, using the *WALS* genus list by Dryer (2013).[19] The second column in Table 2 shows the number of genera in each macroarea and the third column shows the percentage that each macroarea represents of the total number of genera. The remaining columns to the right show the number of languages selected for different sample sizes from each macroarea determined by the percentage. For example, Africa has a total of 74 genera, which equals 14.2 % of the world's total of 521 genera. 14.2 % of the sample languages should thus come from Africa. This means seven African languages in a 50-language PS, 14 in a 100-language PS, and so on for all sample sizes.[20] Once the number of languages has been calculated in the same way for all macroareas, one can start selecting the actual genera and languages.

Table 2: Genera and languages by macroarea with different sample sizes.

	Genera	%	50	100	150	200	300	400	500	600
Africa	74	14.2	7	14	21	28	43	57	71	85
Eurasia	43	8.3	4	8	12	17	25	33	42	50
Southeast Asia & Oceania	66	12.7	6	13	19	25	38	51	64	76
Australia & New Guinea	140	26.9	13	27	40	54	81	108	135	161
North America	92	17.7	9	18	27	35	53	71	89	106
South America	106	20.3	10	20	30	41	61	81	102	122
Total	521	100.0	49	100	149	200	301	401	503	600

Following the principle introduced in Section 4.1, every language in the PS must come from a different genus. To further ensure the genealogical diversity of the PS, the languages should, as far as possible, also come from different families. One starts by randomly selecting a genus from the list of genera in a macroarea, and then selects a language from the selected genus. When selecting the next genus and language, the families from which a language has already been selected are no longer available for selection. One repeats this procedure until the number of

19 The reader will notice that the number of genera in Australia & New Guinea is much higher in the *WALS* list of genera than in the list of genera used in Miestamo (2005), see Table 1. This will affect the percentages to some extent, and shows that the GM method is not immune to the problems of genealogical classification either.

20 Due to rounding effects, some totals in the last row deviate slightly from the predetermined sizes. There are two options to deal with this issue: either accept the deviation and work with a sample of, e.g., 49, 149, 301, 401, or 503 languages, or subtract or add one language from/to the macroarea for which the unrounded number is the closest to allowing one language less or more.

languages needed for that macroarea has been reached. Unless the size of the sample is very small, the number of distinct language families is soon exhausted for some macroareas; once a language has been selected from every family in a macroarea and more languages are still to be selected from that macroarea, the families that are already represented are made available again, and a second round is started. Naturally, the genera from which a language has already been selected are not made available. This procedure is repeated for all macroareas until the desired number of languages has been selected from every macroarea.

To illustrate the family-level stratification, let us take a closer look at Africa. Africa has eight distinct families in the *WALS* classification, and a PS with more than eight languages from Africa has to take a language from more than one genus from some families. Four of the eight families in Africa consist of only one genus, and therefore, if one needs more than 12 African languages in the PS, some families will have to provide three or more languages. With small sample sizes, the requirement of even representation of families will have the effect that no family will be represented by more than one language in some macroareas. This happens when the number of languages to be picked from a given macroarea determined by the percentages in Table 2 is lower than the number of families in the macroarea. In practice one may sometimes have to relax the requirement of even representation of families in cases where adequate sources for the desired genera/languages are not available. It is, however, important that no genus provides more than one language to the PS – the primary unit of genealogical stratification in this methodology is the genus.

If the sample size exceeds the total number of genera (521 in the 2013 edition of *WALS*; Dryer 2013), there will be more than one language from some genera. From each macroarea one then selects a second language from as many genera as needed to reach the number of languages required for the macroarea. Again, one should keep the representation of families even by making sure, to the extent possible, that the genera from which more than one language is selected belong to different families. If the sample size gets large enough, one will exhaust the genera that can supply more than one language to the sample (i.e., genera that are not constituted by only one language). One will then have to start a third round, selecting a third language from a number of genera following the same principles as before. Note, however, that it will be unlikely that sufficient sources could be found for a language from all 521 genera, and therefore, PS sizes approaching, let alone exceeding, the total number of genera remain theoretical in most typological studies. The availability of sources sets an upper limit to the possible size of the PS.

To take areality into account more effectively, the sample may also have a more fine-grained areal stratification within macroareas. As discussed in Section

4.1, the smaller-scale areal classification into 24 areas by Bickel & Nichols (2013) can be used for this purpose. The procedure for this areal stratification is similar to the one just proposed for families: if one language is selected from a given subarea, no other languages are selected from that area until all the subareas of the macroarea have one representative in the PS; if two languages are selected from a given subarea, no other languages are selected from that area until all the subareas of the macroarea have two representatives in the PS, and so on, until the number of languages to be selected from the macroarea is reached. Alternatively, the finer-grained areal stratification could be done by paying attention to geographical distances between languages.

Note finally that, just like CSs in the bottom-up variant of the method, PSs may also be extended. An ES with some genera represented by more than one language may be created for various purposes (see discussion in Section 4.1). The ES adds a bottom-up element to the top-down sampling method. In such a situation, possible quantitative generalizations over the world's languages should be based on the PS, and the ES may be used to address other questions.

4.3 Summary

The main advantages of the GM method are that it combines genealogical and areal stratification, and offers different levels of sampling that can be used for different purposes; it aims to overcome the worst problems of genealogical classifications by relying on the genus level, which is intended as comparable across the world; and it is explicit and can be formalized, thus providing samples commensurable across different studies. Potential problems include the somewhat unclear criteria of determining the limits of genera.

Table 3: Summary of the different levels of sampling and types of samples in the GM method.

Sample size not predetermined:	
Genus Sample (GS)	One language from every genus
Core Sample (CS)	One language from every genus from which a language with usable sources of data is found
Restricted Sample (RS)	Subsample of CS in which the genealogical diversity of each macroarea is equally represented
Extended Sample (ES)	CS plus any additional languages included in the study
Sample size predetermined:	
Primary Sample (PS)	Sample size predetermined, genealogical diversity of every macroarea equally represented
Extended Sample (ES)	PS plus any additional languages included in the study

The different levels of sampling and the samples resulting from the different variants of the GM method are summarized in Table 3.

In practice, when typologists approach different topics, they may use the same database and add data on different domains for essentially the same languages. The flexibility of the overlapping samples in the GM methodology is ideal in this respect. One may have a large CS (or PS), extend it for different purposes, and draw different restricted samples from the CS for different purposes. The CS will of course also differ from research topic to research topic as the usability of data sources is always dependent on the questions one is asking. A language that has sufficient sources for the analysis of, e.g., demonstrative pronouns and can thus be included in the CS when working on demonstrative pronouns, may have to be replaced by another language of the genus when studying, e.g., object marking.

5 Comparing and testing variety sampling methods

The previous two sections have presented and discussed two methods of variety sampling: the DV method and the GM method. In this section, these methods will be tested and compared using a computerized simulation method designed for this purpose. The focus in the simulations is to test existing methods of variety sampling as they have been used in typological work. Possible improvements to the methods will be suggested in the discussion in Section 6, but testing these is a matter of future work.

Recalling that the primary goal of variety sampling is to cover as much of the structural variety shown by the world's languages as possible, it will be interesting to test with real linguistic data how the methods fare in achieving this goal. This can be done by examining how well different sampling methods manage to capture different linguistic types proposed in existing typological studies. In real research settings, variety samples are used in situations where sampling precedes analysis, i.e., one first selects the sample and then proceeds to analyse the data and enter it into one's database.[21] When testing and comparing methods of variety sampling, one has to turn the setting around and try to simulate variety sampling using data available in existing databases such as the *WALS* database. This is what we have done in order to compare the DV and GM

21 In practice sampling and analysis may overlap to some extent since the sample may be affected by the availability of sources.

methods. The results of the simulations are presented in this section. The test procedure consists of generating a set of samples of different sizes for both DV and GM, then linking the languages of each generated sample to linguistic data in the *WALS* database, and finally comparing the linguistic variety displayed by the samples pertaining to the two methods. The choice of *WALS* as the database against which the sampling methods are tested may be criticized,[22] but it has its advantages in that it covers an extensive number of topics in different grammatical domains, includes a high number of languages and data points with a good areal and genealogical coverage, and the data is openly available and freely downloadable; we do not see any obvious alternatives to the *WALS* database.

The GM samples were naturally run on the *WALS* classification, with 521 genera and 2607 languages (i.e., all languages in the 2013 version of *WALS* excluding sign languages as well as pidgins and creoles, since these languages do not constitute genealogical groupings in the relevant sense). DV was run on two different classifications: *Ethnologue* (more specifically, the 15th edition, E15; Gordon (ed.) 2005) and *Glottolog* (GLOT). It could be argued that in order to compare the methods independent of the chosen classification this section should look at DV samples based on the *WALS* classification rather than E15 and GLOT. As the DV method can be applied to any classification that can be represented as a tree, it could in principle be applied to the *WALS* classification as well. There are, however, reasons why it is more interesting to look at the samples based on E15 and GLOT. Firstly, it means comparing the methods as they have been actually used in typological studies. DV samples based on (different editions of) the *Ethnologue* have been used in the typological literature, but no one to our knowledge has based a typological study on a DV sample drawn from the *WALS* classification. Secondly, and more importantly, although the *WALS* classification can be represented as a tree and thus used as a basis for a DV sample, it is not a very fruitful choice to be used with the DV method. It is a very flat classification, consisting of three levels only (or four in the occasional cases where intermediate levels of classification have been taken into account as subfamilies). As the trees have been reduced to these three (or four) levels, a lot of information on the relationships between languages has been omitted from

22 For example, as pointed out by a referee, the criteria used to identify relevant linguistic phenomena across languages may be problematic in some *WALS* chapters; cf. the discussion in Rijkhoff (2009). How exactly this would affect our tests is hard to estimate. In any case, we are comparing one feature at a time and counting mean values but do not make direct comparisons across features, and since it is fair to assume that the criteria of identification used within a given *WALS* chapter are applied consistently to all languages coded for the feature in question, the negative effects of the problem pointed out by the referee do not seem immediately obvious to us.

them. The DV algorithm would not bring much added value when used with such a flat classification, and given that a lot of relevant information is omitted from the trees, counting DVs on the basis of them might lead to the wrong conclusions in many cases. The *WALS* classification simply does not show the full diversity of the families that the DV algorithm aims to reveal.[23] The DV method is primarily designed to be used with trees that have more complexity and depth.

In the E15 classification, language isolates are collapsed into one group that appears as one independent top-level node on a par with independent families. In the GLOT classification, isolates are coded as isolates, but they are not treated as one group but as independent top-level nodes. However, to keep the classifications comparable in the simulations, we have treated isolates in GLOT as they are treated in E15, so there is just one group of language isolates.[24] It can of course be argued that this does not do justice to the idea of DV sampling or to the GLOT classification, but we have chosen to prioritize the comparability of the classifications. It should also be noted that the GLOT sample is constantly updated. The version we use in our simulations dates from 4 February 2014 and differs from the current version in some details.

In order to keep the probabilities under random selection as comparable as possible between the two methods, we restricted the number of languages in E15 and GLOT by excluding all languages not found in *WALS* as well as sign languages and pidgins and creoles, putting this total set at 2,388 for E15 (out of a total of 7,299) and at 2,406 for GLOT (out of a total of 8,038).[25]

23 A DV sample of 521 languages based on the *WALS* classification would have one language per genus, just like the GM method. A DV sample of 215 languages would have one language from each family recognized in the *WALS* classification. With sample sizes smaller than 215, DVs would play a role in determining which families get a representative in the sample, between 215 and 521, they would have a role in determining which families provide more genera to the sample, and with samples larger than 521, they would determine which genera get to be represented by more than one language.

24 If the isolates were treated as independent top-level groups, then all samples smaller than the total number of independent groups (roughly 430 in GLOT) would have one language per group irrespective of the size of the group and in sample sizes between the number of non-isolate families (roughly 240) and all independent groups (430), the method would randomly pick which isolates are included and which ones are not. Below 240 only larger families would be able to make it into the sample and the probability for this would be determined by their DVs. Above 430 the DVs would determine which families would be represented by more than one language.

25 For the necessary linking procedures, use is made of both the *Ethnologue* and *WALS* three-letter coding systems and the Glottocodes used in *Glottolog*.

We included all 144 primary features in the *WALS* database except numbers 139 to 141, which deal with sign languages and writing systems and could not be used in the simulations. In total this means 141 features. In every *WALS* chapter, there is one feature that constitutes the main map in that chapter; these primary features are given the letter A in *WALS*, and accordingly, we have features 1A, 2A, etc. Most chapters have only one feature, but some chapters have more, and then letters B, C, etc. are used. In the simulations, only the primary features (A-features) are included. The reason for this is that many of the non-primary maps contain relatively few languages, which may also be areally restricted, and the features are usually directly dependent on the primary features and would not bring independent datapoints to the simulations. The features we use have anywhere between two (*WALS* Feature 11) and nine (Feature 33) different values, with the exception of Features 143 and 144, which have 17 and 21 values respectively. The frequencies of these values show a great range of variety from single occurences – value 6 for Feature 19 – to 1191 occurences – value 1 for Feature 82. The total number of languages coded for a feature ranges from 112 (Feature 123) to 1,515 (Feature 83) languages. In all, our database contains 76,347 datapoints for the 2,607 languages of the *WALS* database for the version that we used. In terms of data density, this means that there is a value available for around 14.8% of the potential datapoints.

It should perhaps be stressed that the *WALS* database itself is not based on a unified sample built on the basis of a specific method. The database simply contains all languages that the different authors have included in their chapters. The editors provided two *WALS* samples: a set of 100 languages that the authors were asked to include in their chapters if at all possible, and another set of 100 languages that the authors were encouraged to include. In the selection of these samples, the editors had areal and genealogical representativeness as an important guideline. Different authors were able to follow these instructions to different degrees, but nevertheless, the languages of the two *WALS* samples figure in a substantial number of chapters. Many chapters, however, have a much higher number of languages. Despite the fact that the editors provided the two samples, it can be said that the database is a compilation of 144 samples that were selected using different – more or less systematic – methods. The 144 samples have a certain amount of overlap between them due, first and foremost, to the two *WALS* samples, but also to other factors, such as the fact that many maps stem from the same author and may be largely based on the same set of languages, and the fact that the relatively few languages that have been properly described have a much higher chance of ending up in typological samples. In any case, we must assume that the *WALS* database is far from a random selection, and that some general principles of sampling, such as areal and

genealogical diversity, would apply to it, especially since we will be employing the respective subsamples individually.

To be able to generate GM samples, too, we extended the computer programme designed to create DV samples as discussed in Rijkhoff et al. (1993) in several ways. Firstly, we built in the GM method as presented in Section 4 such that we can now also automatically create Primary Samples of any predetermined size.[26] In this computerized version, any PS of size *n* starts with a genus sample containing precisely one language per genus in the classification, with a total of 521 languages in the current version. This sample is then either reduced or expanded, on the basis of the proportional numbers of languages assigned to the respective macroareas. Reduction takes place by deselecting genera such that, per linguistic area and iteratively, a genus is taken away from the largest family in terms of genera still present, thus maintaining the highest number of different families represented by at least one genus as long as possible. The genus to be eliminated from a family is always the one containing the smallest number of languages in the classification. When the sample size for an area is smaller than the number of families that belong to it, random selection takes place to eliminate families from the sample. For sample sizes above 521, extra languages are assigned to genera such that, per area and iteratively, an extra language is added to the family with the lowest number of genera with more than one language assigned to it. Within the selected family, the genus with the lowest number of languages assigned to it will get the extra language. When all genera within an area have been assigned two languages, and the total sample size for an area has not been reached yet, a third language will be added, provided that a genus contains enough languages. In all cases, when more than one genus qualifies for either elimination or extension, random selection takes place.

For either of the sampling methods we generated a number of samples for a range of representative sample sizes. The set of sample sizes was established at 50, 100, 150, ..., 900. We think that this range covers the sizes of the vast majority of samples in the typological literature, as well as over 90 % of the samples in the *WALS* database. For any tuple <Sample Strategy, Sample Size, *WALS* Feature> the programme may be instructed to make any number of simulations, i.e., resamplings for the same tuple. We decided to run 10 simulations for each tuple; extensive experimentation with different numbers of

26 Note that if we were simulating the GM method independently, we would draw RSs from individual *WALS* maps as demonstrated in Section 4.1. But since the DV method cannot be simulated this way and the simulations of the two methods have to be kept similar in order for them to be comparable, our simulations will draw PSs.

simulations had shown us that the results would remain similar irrespective of whether we would run 10 or 100 simulations per tuple. 10 simulations, 18 sample sizes, and 141 *WALS* features give us 25,380 samples per method. For each of these samples, we made a random draw of the required number of languages from the genealogical groupings as indicated for the respective samples, irrespective of the presence or absence of a value for any of the features in the *WALS* database. In case a given language that was selected did not have a value for the *WALS* feature, it was assigned the value of zero, and thus did not contribute to the variety in that sample.

Per resulting selection we determined two quantities, Saturation and Completeness, which are designed to express the relative success rate of each of the methods for a sample of that size.[27] The SATURATION (SAT) of a sample for some feature is the proportion of values, out of the maximum number of possible values, found in the sample for that feature, i.e., the number of different values present in the sample divided by the total number of different values present in the database for that feature. Having extracted ten samples/draws per each feature and sample size (and method), per each feature we use the mean of the ten individual Saturation values for that feature in the assessments of the significance of the differences between the two methods later below. Complete Saturation (SAT = 1.0) for an individual sample means that all the different values are attested at least once in that sample. The overall COMPLETENESS (COMP) for some feature and sample size is the proportion of draws in which complete Saturation was reached. For instance, a value of .5 would mean that, out of 10 draws, in 5 draws all values for the corresponding feature were attested, i.e., in 5 draws the SAT value for the feature in question was 1.0. In short, for a sample of a certain size, SAT represents the overall capacity of a sampling method to find as many of the different values for some feature as possible, and COMP represents its capacity to find ALL of these values as often as possible, representing the chance (as a historical proportion) that all values will be found. It is on the basis of these two quantities that we will compare the DV and GM methods.

Table 4 shows the mean Saturation and Completeness values for GM and DV (in both E15 and GLOT) over 10 samples for each of the 141 features and all sample sizes running from 50 to 900. As can be seen, GM performs slightly better (on average) than DV on both Saturation and Completeness when DV is run on the GLOT classification, while the DV scores run on E15 are, in turn, slightly better than the GM scores. Note that the DV scores may be slightly

27 We experimented with a number of different possible measures of variety. Saturation and Completeness seem to us to be the most adequate for the present purpose.

skewed in a positive sense by the fact that the number of languages included in the DV simulations is somewhat lower than the number of languages in *WALS*. In any case, our conclusion is that the overall means show roughly equal performance for DV and GM. We will come back to the differences between the classifications later.

Table 4: Mean Saturation and Completeness over all sample sizes: GM and DV (E15/GLOT).

	GM (WALS) *n* = 2607	DV (E15) *n* = 2388	DV (GLOT) *n* = 2406
SAT	0.905	0.921	0.899
COMP	0.690	0.716	0.662

These global numbers do not, however, tell us anything about possible differences between the methods depending on sample size or individual features. We will now look at the effects of sample size. Figure 1 shows the mean Saturation values for DV (E15 and GLOT) and GM for the different sample sizes. Figure 2 gives the Completeness values per method and sample size.

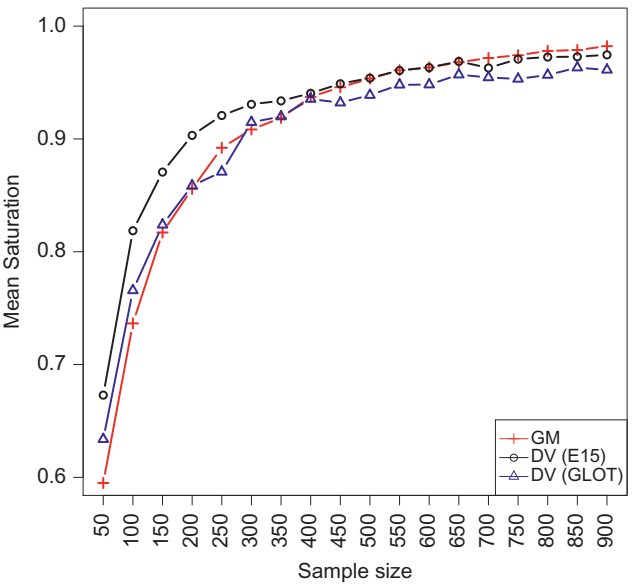

Figure 1: Mean Saturation and sample size: GM and DV (E15/GLOT).

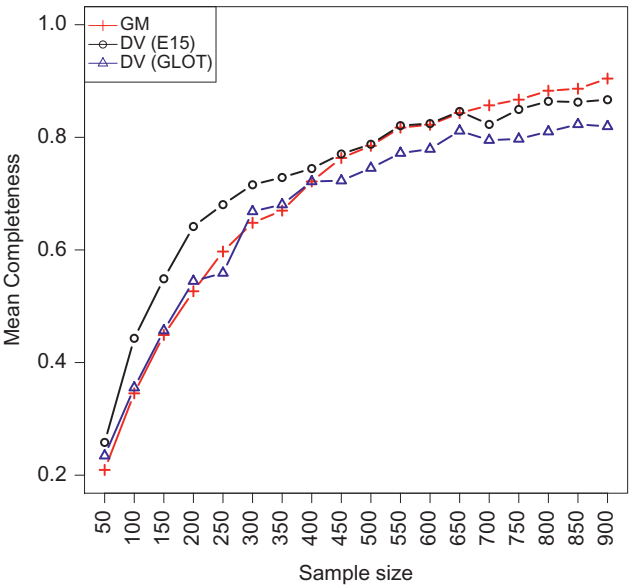

Figure 2: Mean Completeness and sample size: GM and DV (E15/GLOT).

As can be seen from Figures 1 and 2, DV run on E15 appears to perform better than DV run on GLOT with respect to both Saturation and Completeness with all sample sizes. Comparing the two methods, we observe that with lower sample sizes the mean performance of GM is roughly equal to DV run on GLOT, but with larger sample sizes GM improves and reaches DV run on E15, and outperforms it with the largest sample sizes. We may note here that the dividing point corresponds roughly to the number of genera (521) in the *WALS* classification. Importantly, the differences between the distributions of the mean Saturation values (calculated per each feature out of individual Saturation values for the 10 samples/draws; resulting in 139 Saturation values for each sample size)[28] between DV (E15) and GM are significant,[29] and in favour of DV, only for the smaller sample sizes up to 350; however, the situation reverses so that GM performs significantly better than DV (E15) at the largest sample sizes 700 and 800–900. The differences with mean Saturation values between DV (GLOT) vs. GM, are

28 In these calculations of statistical significance, the two features with more than ten values, namely 143A and 144A, were excluded for practical reasons.

29 All assessments of statistical significance of distributions of Saturation and Completeness measures between various variants of the methods are based on non-parametric Wilcoxon's one-tailed paired tests per each sample size, with a critical $p < .05$ for significance.

significantly in favour of DV only for the smallest sample sizes 50 and 100, and again the performance reverses, GM being significantly better with all the larger sample sizes from 450 upwards.

As for mean Completeness values, the differences between DV (E15) and GM are significantly in favour of DV for the smaller sample sizes of 50 through 350, while again GM performs significantly better for the largest sample sizes, from 700 through 900. The differences between DV (GLOT) vs. GM are significantly in favour of DV only for the smallest sample size of 50, but significantly in favour of GM for the all the larger sample sizes from 450 upwards.

We have seen that both methods show Saturation and Completeness values close to the maximum (1.0). Both methods thus seem to work relatively well for the purpose of capturing crosslinguistic variety. To put this impression to the test, we ran the corresponding simulations without any (stratified) sampling method, i.e., drawing the same number of random samples without any stratification for all 18 sample sizes and all 141 *WALS* features. Table 5 and Figures 3 and 4 compare the methods against random sampling. The latter was done on the smallest of the language sets, i.e., the one used for E15. Arguably, this is a "best case" analysis, and the results for random sampling would be even lower if we had used the larger *WALS* sets.

Table 5: Mean Saturation and Completeness over all sample sizes: GM, DV (E15/GLOT), and Random (E15).

	GM (WALS) $n = 2607$	DV (E15) $n = 2388$	DV (GLOT) $n = 2406$	Random $n = 2388$
SAT	0.905	0.921	0.899	0.772
COMP	0.690	0.716	0.662	0.428

As can be seen in Figure 3, the mean Saturation values of the simulations based on the two variety sampling methods are considerably higher than those for the random samples of the same size; the same can be observed for Completeness in Figure 4. Crucially, the distributions of both the Saturation and Completeness values for both the DV and GM methods are significantly better than those for random samples for any sample size (Wilcoxon paired one-tailed tests per each sample size, $p < .05$).

On the basis of the results presented in this section, we can draw the following conclusions. Using a method of variety sampling, either GM or DV, significantly increases the chances of capturing the linguistic variety shown by

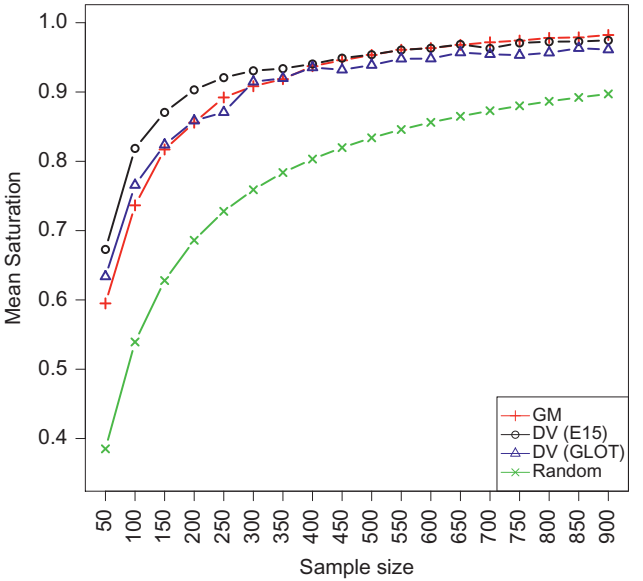

Figure 3: Mean Saturation and sample size: GM, DV (E15/GLOT), and Random.

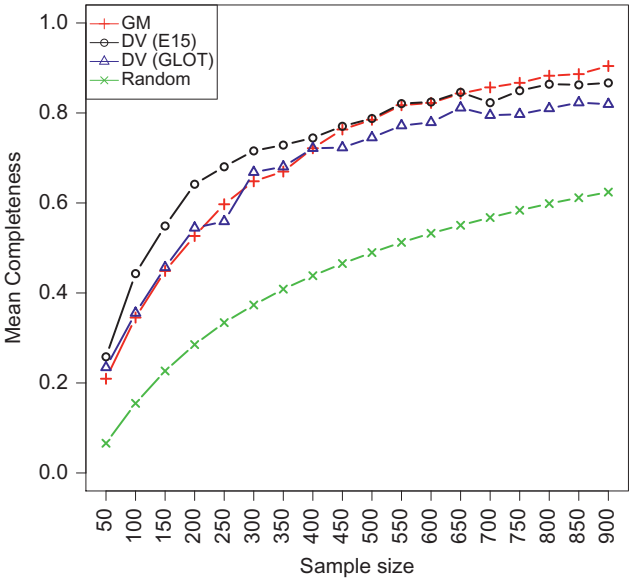

Figure 4: Mean Completeness and sample size: GM, DV (E15/GLOT), and Random.

the world's languages, as compared to working on a random sample without stratification. This holds for any sample size. It is worth pointing out that many typological studies are based on convenience samples with areal and genealogical biases, while the tests in this section were made against unstratified random sampling. A comparison against convenience sampling would be likely to show even clearer differences in favour of our two variety sampling methods. A more detailed discussion of the results will follow in Section 6.

6 Discussion and conclusions

In this section, we will discuss the results from different angles, suggesting also directions for future research. The clearest result of the simulations presented in Section 5 is that the variety sampling methods perform much better in finding the crosslinguistic variety in the data than random sampling without stratification does. This result is in line with what was said about the reasons for stratification in Section 1: stratification guarantees that different language groupings are represented in the sample to the best possible degree relative to sample size and that the sample languages are as independent of each other as possible both genealogically and areally; increasing the genealogical and areal variety of the sample in this way contributes to increasing its linguistic variety. We have thus shown, with concrete evidence, that using a variety sampling method is to be preferred over NOT using such a method. Once we have established that a variety sampling method is needed, the next question is what kind of method should be chosen. This is a question that our comparison between the two methods aimed to find answers for, but the answers we get are not as clear as the first result concerning the use vs. non-use of a method.

The results do not show categorical differences between the two methods, i.e., either one always performing significantly better or worse than the other, and the choice of classification to be used with the DV method (E15 vs. GLOT) seems to be at least as important as the choice of method (DV vs. GM). DV run on E15 gives better results than DV run on GLOT with all sample sizes. With smaller sample sizes the mean GM results are roughly equal to the mean results of DV (GLOT) and with larger sample sizes they are roughly equal to those of DV (E15); however, the overall distributions of the Saturation and Completeness values with the smaller sample sizes are significantly better for both DV (E15) and DV (GLOT) over GM, while with the larger sample sizes the distributions of both values are reversed, being significantly better for GM over both DV (E15) and DV (GLOT). The results raise a number of questions:

(i) Why does a method that includes areal stratification (GM) not perform better throughout all sample sizes than a method without areal stratification (DV)?
(ii) Why does sample size have the observed effect on the performance of the methods?
(iii) Why does DV run on E15 perform better than DV run on GLOT?

Since the focus of this article is on comparing the two methods, we will concentrate on question (i), but pay attention to the other two questions as well, to the extent that they may help us understand the difference between the two methods.

Given that both genealogical and areal biases can be harmful for the goals of variety sampling, it is somewhat unclear why adding areal stratification does not simply improve the results.[30] Surely, we do not want to draw the conclusion that areal stratification need not be used in variety sampling. The differences between the methods lie both in the presence vs. absence of areal stratification and in different principles of genealogical stratification. Given that the performance of the GM method first improves with respect to DV with sample sizes that are roughly equal to the total number of genera (521), and then surpasses DV with even larger sample sizes, our first suspicion would be that the number of languages to be sampled from each macroarea in the GM method is not optimal, leading to harmful biases in samples in which genera are eliminated on the basis of these numbers. In its present form, the number of languages to be sampled from each macroarea is determined by the number of genera in each macroarea. This ensures that the genealogical diversity of each macroarea is equally well represented in the sample, but on the other hand, it may lead to areal bias in the sense that an area like New Guinea, which has a lot of genealogical diversity, is very strongly presented in the sample and contact-based similarities between the languages in this area may get overrepresented.

In order to see how the different macroareas are represented in the other two approaches, DV (E15) and DV (GLOT), we examined the macroareal distribution of languages in DV samples of different sizes (50 to 900) generated on E15 and GLOT. Tables 6 and 7 show the distributions for DV (E15) and DV (GLOT), respectively, and Table 8 gives the corresponding numbers for GM for comparison.

30 One reviewer suggests that the reason why adding areal stratification does not noticeably improve the results is that since language families are mostly compact, both types of classification coincide largely. However, as we found large differences in language density in the respective macroareas between the different methods and classifications (cf. Tables 6 through 8), this is probably not the case.

Table 6: Languages per macroarea and sample size: DV (E15).

Area	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Africa	4	4	20	35	47	59	73	84	97	109	120	132	143	156	166	179	189	200
Eurasia	8	11	15	19	26	31	37	42	47	52	58	64	68	76	81	87	93	97
Southeast Asia & Oceania	4	6	27	44	58	72	86	99	114	128	141	156	168	182	196	208	222	235
Australia & New Guinea	10	15	23	33	43	52	64	70	83	90	101	110	120	127	137	146	155	167
North America	11	27	27	29	32	38	39	49	50	57	61	65	74	77	83	88	95	101
South America	13	35	36	38	42	46	49	54	57	62	67	71	75	80	85	90	94	98

Table 7: Languages per macroarea and sample size: DV (GLOT).

Area	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Africa	7	18	17	31	38	56	73	87	100	115	128	142	155	167	180	191	206	219
Eurasia	6	9	12	16	16	23	28	36	44	50	56	64	71	78	81	91	95	103
Southeast Asia & Oceania	4	4	5	5	10	28	50	67	87	101	116	130	144	160	174	188	203	215
Australia & New Guinea	13	32	65	77	100	106	111	116	121	128	137	144	153	163	172	181	189	198
North America	8	19	24	33	40	40	40	43	44	48	50	53	56	58	63	66	68	72
South America	12	16	27	37	44	45	46	49	52	56	61	65	69	72	78	81	87	91

Table 8: Languages per macroarea and sample size: GM.

Area	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Africa	7	14	21	28	36	43	50	57	64	71	78	85	92	99	107	114	121	128
Eurasia	4	8	12	17	21	25	29	33	37	41	45	50	54	58	62	66	70	74
Southeast Asia & Oceania	6	13	19	25	32	38	44	51	57	64	70	76	82	89	95	101	108	114
Australia & New Guinea	13	27	40	54	67	81	94	108	121	135	148	161	175	188	202	215	228	242
North America	9	18	27	35	44	53	62	71	79	89	97	106	115	124	132	141	150	159
South America	10	20	30	41	51	61	71	81	92	102	112	122	132	142	153	163	173	183

The original samples on which the numbers in Tables 6 and 7 are based are given in the Appendix.

An immediate general observation is that the representation of macroareas grows linearly in the GM samples as sample size increases, but in DV samples (E15 and GLOT) the increase is more haphazard. Looking at the similarities and differences in the numbers in more detail, we can first compare DV (E15) in Table 6 with GM in Table 8 and observe that Africa, Eurasia, and Southeast Asia & Oceania get a much higher representation in DV (E15) except for the smallest sample sizes whereas for Australia & New Guinea, North America, and South America the pattern is the opposite, showing a much lower number of languages for these areas for all but the smallest sample sizes. As to the smallest sample sizes, Africa and Southeast Asia & Oceania have a lower number of languages for DV (E15) than GM and the Americas have a higher number with sample size 100. The comparison between DV (GLOT) and GM gives a similar overall picture, but the point where the higher representation for the first three macroareas and the lower representation for the latter three start is somewhat higher than with DV (E15); a further difference is seen in Southeast Asia & Oceania that shows clear underrepresentation with the lowest sample sizes up to 300 and Australia & New Guinea that shows overrepresentation with sample sizes 150 to 350. Clear differences are thus observed in how the different macroareas are represented in GM vs. DV (E15) and DV (GLOT). However, these differences do not seem to map very well to the differences in the simulation results. The results of the simulations are not very big and the biggest differences are between the DV simulations run on different classifications, GM falling in between these. If the macroareal distributions were to explain the differences, they should show differences between DV (E15) and DV (GLOT) rather than between DV samples and GM samples. Consequently, it seems that we cannot get very far in explaining the results by the different macroareal distributions.

If it is assumed that the principle of proportional representation of the genealogical diversity of macroareas as defined in the GM method is a valid desideratum for variety sampling, the deviations from these ideals in the examined DV samples should show as worse results in our simulations in Section 5, but this is not the case. Given that no clear differences in the results were found despite the different macroareal representations, we could conclude that areal stratification does not play a big role in explaining the results, and focus on the genealogical stratification method and the underlying genealogical classification in order to try to understand the different results. The differences in the principles of genealogical stratification used in DV and GM have been discussed at length above, but since the clearest difference in the

simulation results is between the two DV simulations and not between DV and GM, we will not come back to the differences between the methods here; instead we will focus on the differences between the E15 and GLOT classifications. The main difference in the classifications has to do with conservativeness in assuming higher-level connections. Unlike the classification by Ruhlen (1991), neither one of the classifications is a lumping classification proposing large macrofamilies without proper evidence, but the GLOT classification is even more cautious in this respect and requires more rigorous evidence based on up-to-date scholarship to assume genealogical relatedness.[31] Consequently, the GLOT classification has a higher number of independent top nodes whereas in E15 the families tend to be somewhat larger and more diverse. For the DV method, this has the consequence that large and diverse families get somewhat better represented when DV is run on E15 than on GLOT, but this difference should only show with larger sample sizes and since DV (E15) performs better than DV (GLOT) with all sample sizes, this overall difference in the classifications does not explain the results. Another difference we could consider in this context is the extent to which the DV samples drawn on the basis of the two classifications would differ in their adherence to the GM principle of one language per genus – with larger sample sizes some large genera such as Bantoid or Oceanic would get represented by more languages than the GM principles would allow, but there is no obvious way in which the two classifications would differ in this respect. A detailed look at how DV samples drawn from E15 and GLOT differ in terms of the representation of different families and lower-level groups could perhaps give insight into why the simulations based on the two classifications differ, but since the focus of this article is on the comparison of methods, not classifications, we will not delve deeper into this question.

In the preceding discussion, we have tried to find an explanation for the differences in the simulation results for the two methods. No clear explanations were found, but, as we also noted, the differences between the results of the methods are not very great nor uniformly in favour of one over the other, so there is not so much to find explanations for. Turning the question around, we can come back to our original perspective asking why a method combining areal stratification with genealogical stratification did not give better results than the purely genealogical method without areal stratification. We have already shown that using either one of the methods yields much better results than non-stratified random sampling. But if we want to

31 See http://glottolog.org/glottolog/glottologinformation for discussion.

improve our results and find an even better method of variety sampling, then we should try to find the optimal combination of the right kind of genealogical and areal stratification. We may ask whether the reason for GM not performing better than DV lies in the genealogical or in the areal stratification method (or both).

Starting with areal stratification, the easiest way to adjust the GM method would be to use alternative numbers of languages per macroarea. As pointed out above, determining the number of languages to be sampled from each macroarea purely on the basis of the genealogical diversity (number of genera) of the macroareas, might lead to overrepresentation of areas such as New Guinea where the number of genera is high but there has been relatively much contact between the languages, levelling out typological differences to some extent so that the area is not typologically as diverse as the number of genera might lead one to think. As discussed in Section 2.2, Dahl (2008) examined the typological diversity of continents in the *WALS* database and suggested that this should determine the number of languages to be sampled from them. Dahl's numbers for a 101-language sample are: Africa 16, Europe 8, Asia 17, Oceania 3, New Guinea 9, Australia 9, North America 19, and South America 20. A direct comparison of Dahl's numbers with the numbers in Tables 6 through 8 cannot be made since the macroareal divisions do not match Dahl's continents as regards Eurasia and Southeast Asia & Oceania. For the remaining macroareas, we may note that the GM numbers in Table 8 are higher for Australia & New Guinea, but that the proportions are roughly equal for Africa, North America, and South America. As regards the DV samples in Tables 6 and 7, it can be observed that North America and South America get a much lower representation with the larger sample sizes as compared to the GM numbers and to Dahl's recommendations. The underrepresentation of these typologically diverse areas may partly explain why the performance of the DV method is weaker with the larger sample sizes. One way of adjusting the GM method would be to adopt this principle and let the number of languages to be sampled from each macroarea be determined by the typological diversity of the macroarea. Adopting Dahl's numbers as such in simulations of GM sampling would be problematic for two reasons: Dahl's numbers are not directly applicable to the GM method because the continental divisions he uses are different from the GM macroareas, and more importantly, there would be a potential danger of circularity in testing the GM method's ability to find typological variety in the *WALS* database using as a basis for stratification numbers that are determined by typological diversity in the same database. The principle introduced by Dahl is worth testing in future simulations if

typological diversity scores for the macroareas can be established from an independent source.

Further possible ways to search for improvements to the areal stratification method in future work include adopting alternative macroareal divisions, e.g., the ones proposed by Bickel & Nichols (2013) and by Hammarström & Donohue (2014); see Section 2.1 for a brief discussion of these. Furthermore, it is possible that macroareal stratification is too coarse for the purposes of variety sampling, and positive effects of areal stratification only start to show up when areal stratification is done at a more fine-grained level. To test this, finer areal divisions within macroareas could be implemented as a further level of areal stratification as suggested in Section 4.2. Finally, areal stratification within macroareas could also be based on geographical distances between languages,[32] i.e., by building in some generally acknowledged distance measure into the simulations, or by experimentally finding one.

Searching for the optimal method of variety sampling in future work should also include experimenting with different kinds of genealogical stratification. One could try combining the DV method of genealogical classification with different types of areal stratification to see whether that would lead to better results than GM-style genus-based stratification. One thing that should certainly be done in future simulations of DV sampling would be to treat isolates as independent families; as was mentioned in Section 5, isolates were treated as one group in the DV simulations, due to the structure of the E15 classification. Given the high number of isolates in the GLOT classification, this might lead to overrepresentation of areas that contain high numbers of isolates (e.g., New Guinea), but adding areal stratification could remove the potential negative effect.

We have not presented results separately for different *WALS* features, but it could also be worth examining which features are best captured by which sampling method. This would deepen our understanding of the strengths and weaknesses of each method – one method may give better SAT and/or COMP values with certain kinds of feature value distributions, while another method may work better with other kinds of distributions. We have experimented with testing this, too, and observed that DV gives better results with only a handful of *WALS* features, while a couple of other features are better captured by GM. But most features show no significant difference between the two methods. Comparing the methods feature by feature is beyond the scope of this article, but could be an interesting additional element in future simulations.

32 We thank an anonymous referee for pointing this out.

A central contribution of this article has been the development of a method for evaluating – simulating and comparing – the performance of variety sampling methods. This evaluation method was used to compare two sampling methods proposed in prior literature. The results of the comparison are interesting as such and the lessons learnt from the exercise can be used in future research to develop new and better methods of variety sampling. The preceding discussion has proposed several avenues to test in the quest for the optimal sampling method.

Acknowledgements: We are grateful to Kaius Sinnemäki and the three anonymous reviewers for their valuable comments on earlier versions of this article. We also wish to thank the audience at the ALT 9 conference in Hong Kong in 2011.

Abbreviations: COMP = completeness; CS = core sample; DV = diversity value; E13 = *Ethnologue*, 13th edn. (Grimes (ed.) 1996); *E*15 = *Ethnologue*, 15th edn. (Gordon (ed.) 2005); *E*18 = *Ethnologue*, 18th edn. (Lewis et al. (eds.) 2015); ES = extended sample; *GLOT* = *Glottolog* (Hammarström et al. 2015); GM = genus-macroarea; GS = genus sample; PS = primary sample; RS = restricted sample; SAT = saturation; *WALS* = *World atlas of language structures* (Haspelmath et al. (eds.) 2005; Dryer & Haspelmath (eds.) 2013).

Appendix: DV samples discussed in Section 6

Table A-1: Numbers of languages included in DV samples based on E15.

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Africa	**4**	**4**	**20**	**35**	**47**	**59**	**73**	**84**	**97**	**109**	**120**	**132**	**143**	**156**	**166**	**179**	**189**	**200**
Afro-Asiatic	1	1	4	7	9	12	15	18	21	23	26	29	31	34	36	39	41	44
Khoisan	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Niger-Congo	1	1	12	21	29	36	43	49	56	63	69	76	82	89	95	102	108	114
Nilo-Saharan	1	1	3	6	8	10	13	15	18	20	22	24	26	29	31	33	35	37
Eurasia	**8**	**11**	**15**	**19**	**26**	**31**	**37**	**42**	**47**	**52**	**58**	**64**	**68**	**76**	**81**	**87**	**93**	**97**
Altaic	1	1	1	2	2	3	4	5	5	6	7	7	8	9	9	10	11	12
Andamanese	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Basque	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Chukotko-Kamchatkan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Dravidian	1	1	1	1	2	3	3	4	4	5	5	6	6	7	8	8	9	9
Indo-European	1	1	4	7	11	14	17	20	23	26	29	32	34	38	40	43	46	48
Japanese	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Kartvelian	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
North Caucasian	1	1	1	1	2	2	3	3	4	4	5	5	5	6	6	7	8	8
Uralic	1	1	1	1	2	2	3	3	4	4	5	6	6	7	7	8	8	9
Yeniseian	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yukaghir	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(*continued*)

Table A-1: (*continued*)

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Southeast Asia & Oceania	**4**	**6**	**27**	**44**	**58**	**72**	**86**	**99**	**114**	**128**	**141**	**156**	**168**	**182**	**196**	**208**	**222**	**235**
Austro-Asiatic	1	1	3	4	6	8	9	11	13	15	17	19	20	22	24	25	27	28
Austronesian	1	1	18	31	40	50	59	67	77	86	94	103	112	121	130	138	147	156
Cant	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hmong-Mien	0	1	1	1	1	1	1	2	2	2	3	3	3	3	4	4	4	5
Sino-Tibetan	1	1	3	6	8	10	13	14	17	19	21	24	26	28	30	32	34	36
Tai-Kadai	1	1	1	1	2	2	3	4	4	5	5	6	6	7	7	8	9	9
Australia & New Guinea	**10**	**15**	**23**	**33**	**43**	**52**	**64**	**70**	**83**	**90**	**101**	**110**	**120**	**127**	**137**	**146**	**155**	**167**
Amto-Musan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Australian	1	1	4	7	10	12	16	18	22	24	27	29	32	34	37	39	42	45
Bayono-Awbono	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
East Bird's Head	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
East Papuan	1	1	1	1	1	2	2	2	3	3	3	4	4	4	4	5	5	6
Geelvink Bay	0	1	1	1	1	1	1	2	2	2	3	3	3	3	4	4	4	4
Harakmbet	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kwomtari-Baibai	0	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
Left May	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lower Mamberamo	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sepik-Ramu	1	1	1	2	3	4	5	5	6	7	8	9	10	11	12	12	13	14
Sko	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Torricelli	1	1	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	9
Trans-New Guinea	1	1	6	12	17	22	27	30	36	39	44	48	52	56	60	65	69	73
West Papuan	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5

North America	11	27	27	29	32	38	39	49	50	57	61	65	74	77	83	88	95	101
Algic	1	1	1	1	2	2	2	3	3	4	4	5	5	5	6	6	7	7
Caddoan	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Chimakuan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Chumash	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Coahuiltecan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eskimo-Aleut	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Gulf	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hokan	1	1	1	2	2	3	3	4	4	5	6	6	7	7	8	9	9	10
Huavean	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Iroquoian	1	1	1	1	1	1	1	2	2	2	3	3	3	3	4	4	4	4
Keres	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kiowa-Tanoan	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Mayan	1	1	1	1	2	2	3	4	4	5	5	6	6	7	7	8	9	9
Mixe-Zoque	0	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3
Muskogean	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Na-Dene	1	1	1	1	1	2	2	3	3	4	4	4	5	5	6	6	7	7
Oto-Manguean	1	1	1	1	2	3	3	4	4	5	5	6	7	7	8	8	9	10
Penutian	0	1	1	2	2	3	3	4	5	5	6	7	7	8	8	9	10	10
Salishan	0	1	1	1	1	2	2	3	3	3	4	4	4	5	5	6	6	7
Siouan	0	1	1	1	1	1	1	2	2	2	2	2	3	3	3	3	4	4
Subtiaba-Tlapanec	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tarascan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(*continued*)

Table A-1: (*continued*)

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Totonacan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Uto-Aztecan	0	1	1	1	1	2	2	3	3	4	4	4	5	5	6	6	7	7
Wakashan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Witotoan	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Yuki	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
South America	**13**	**35**	**36**	**38**	**42**	**46**	**49**	**54**	**57**	**62**	**67**	**71**	**75**	**80**	**85**	**90**	**94**	**98**
Alacalufan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Arauan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Araucanian	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Arawakan	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Arutani-Sape	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Aymaran	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Barbacoan	0	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3	3
Cahuapanan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Carib	1	1	1	1	1	2	2	3	3	3	4	4	4	5	5	6	6	7
Chapacura-Wanham	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Chibchan	1	1	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8
Choco	1	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	6	6
Chon	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Guahiban	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hibito-Cholon	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Jivaroan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Katukinan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Lule-Vilela	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Macro-Ge	1	1	1	2	3	4	4	5	6	7	8	9	9	10	11	12	13	13				
Maku	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Mascoian	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Mataco-Guaicuru	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
Misumalpan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Mura	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Nambiquaran	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Panoan	1	1	1	1	2	2	2	3	3	4	4	5	5	6	6	6	7	7				
Peba-Yaguan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Quechuan	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2				
Salivan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Tacanan	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2				
Tucanoan	1	1	1	1	1	2	2	3	3	3	4	4	4	5	5	6	6	6				
Tupi	0	1	1	2	3	3	4	5	5	6	7	8	8	9	10	11	11	12				
Uru-Chipaya	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Yanomam	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Zamucoan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Zaparoan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Language ilsolate	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Unclassified	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				

Table A-2: Numbers of languages included in DV samples based on GLOT.

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Africa	**7**	**18**	**17**	**31**	**38**	**56**	**73**	**87**	**100**	**115**	**128**	**142**	**155**	**167**	**180**	**191**	**206**	**219**
Afro-Asiatic	1	1	1	1	1	5	7	9	12	16	19	23	25	29	32	34	38	40
Atlantic-Congo	1	1	1	1	6	20	35	45	55	63	71	80	88	96	104	111	119	126
Central Sudanic	1	1	1	1	1	1	1	2	2	3	4	4	5	5	6	6	7	7
Daju	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Dizoid	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Dogon	0	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3
Eastern Jebel	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Furan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Gonga-Gimojan	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3
Heiban	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Ijoid	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kadugli-Krongo	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Katla-Tima	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Khoe-Kwadi	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
Koman	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kresh-Aja	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kuliak	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kxa	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Maban	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mande	1	1	1	1	1	1	1	2	2	3	3	3	4	4	5	5	6	6
Mao	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Narrow Talodi	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Nilotic	0	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	4	4	4	4
Nubian	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Nyimang	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Rashad	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Saharan	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Songhay	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
South Omotic	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Surmic	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Tama	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Temein	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tuu	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eurasia	**6**	**9**	**12**	**16**	**16**	**23**	**28**	**36**	**44**	**50**	**56**	**64**	**71**	**78**	**81**	**91**	**95**	**103**					
Abkhaz-Adyge	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
Chukotko-Kamchatkan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
Dravidian	0	1	1	1	1	1	1	1	2	2	3	3	4	4	4	5	5	6					
Great Andamanese	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
Hurro-Urartian	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
Indo-European	1	1	1	1	1	8	13	20	27	31	36	41	46	52	55	60	64	68					
Japonic	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2					
Jarawa-Onge	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
Kartvelian	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
Mongolic	1	0	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3					
Nakh-Daghestanian	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	3	3	3					
Tungusic	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2					

(continued)

Table A-2: (*continued*)

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Turkic	0	1	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4
Uralic	1	1	1	1	1	1	1	2	2	3	3	4	4	5	5	6	6	7
Yeniseian	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yukaghir	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Southeast Asia & Oceania	**4**	**4**	**5**	**5**	**10**	**28**	**50**	**67**	**87**	**101**	**116**	**130**	**144**	**160**	**174**	**188**	**203**	**215**
Austroasiatic	1	1	1	1	1	3	4	5	7	9	10	12	13	16	18	19	22	23
Austronesian	1	1	1	1	5	15	29	38	48	55	62	69	76	84	90	97	104	110
Hmong-Mien	0	0	1	1	1	1	1	1	1	1	2	2	2	2	3	3	3	3
Sino-Tibetan	1	1	1	1	2	8	14	21	28	33	38	43	48	53	57	62	67	71
Tai-Kadai	1	1	1	1	1	1	2	2	3	3	4	4	5	5	6	7	7	8
Australia & New Guinea	**13**	**32**	**65**	**77**	**100**	**106**	**111**	**116**	**121**	**128**	**137**	**144**	**153**	**163**	**172**	**181**	**189**	**198**
Alor-Pantar	1	0	1	1	1	1	1	1	1	1	2	2	2	2	3	3	3	4
Amto-Musan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Angan	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3
Arafundi	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Awin-Pa	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Baibai-Fas	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Baining	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bayono-Awbono	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Border	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bosavi	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bulaka River	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Bunaban	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Dagan	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Doso-Turumsa	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
East Bird's Head	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
East Kutubu	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
East Strickland	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
East Timor-Bunaq	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eastern Daly	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eastern Trans-Fly	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Garrwan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Geelvink Bay	0	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	2	3
Giimbiyu	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Goilalan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Gunwinyguan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
Harakmbut	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hatam-Mansim	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Inanwatan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Inland Gulf	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Iwaidjan Proper	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Jarrakan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kaure-Narau	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kayagar	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kiwaian	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Koiarian	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kolopom	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Konda-Yahadian	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kwalean	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kwerbic	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(*continued*)

Table A-2: (*continued*)

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Kwomtari	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lakes Plain	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
Left May	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lepki-Murkim	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Limilngan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lower Sepik-Ramu	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4
Mailuan	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
Mairasi	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mangarrayi-Maran	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Maningrida	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Manubaran	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Marind	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Marrku-Wurrugu	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Maybrat	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mirndi	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mombum	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mongol-Langam	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Monumbo	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Morehead-Wasur	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Namla-Tofanma	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ndu	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Nimboran	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
North Bougainville	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
North Halmahera	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2

North-Eastern Tasmanian	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Northern Daly	0	0	0	0	1	1	1	8	10	13	16	19	21	25	27	30	33	34						
Nuclear Eleman	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Nuclear Torricelli	1	1	1	1	1	1	2	3	3	4	5	5	6	7	8	8	9	10						
Nuclear Trans New Guinea	1	1	1	1	1	4	6	8	10	13	16	19	21	25	27	30	33	34						
Nyulnyulan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Pahoturi	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Pama-Nyungan	1	1	1	1	1	4	6	8	10	12	15	18	21	24	26	29	32	33						
Pauwasi	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Piawi	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Senagi	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Sentani	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Sepik	1	0	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	5						
Sko	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Somahai	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
South Bird's Head	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
South Bougainville	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
South-Eastern Tasmanian	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Southern Daly	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Suki-Gogodala	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Tangkic	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Taulil-Butam	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Teberan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Tirio	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Tor-Orya	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3						
Turama-Kikori	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Umbugarla	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
Walio	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						

(*continued*)

Table A-2: (*continued*)

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
West Bird's Head	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
West Bomberai	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Western Daly	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Western Tasmanian	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Worrorran	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yangmanic	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yareban	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yawa	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yuat-Maramba	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
North America	**8**	**19**	**24**	**33**	**40**	**40**	**40**	**43**	**44**	**48**	**50**	**53**	**56**	**58**	**63**	**66**	**68**	**72**
Algic	1	1	1	1	1	1	1	2	2	3	3	4	4	5	5	6	6	7
Athapaskan-Eyak-Tlingit	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4
Caddoan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Chimakuan	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Chinookan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Chumashan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Cochimi-Yuman	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Coosan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eskimo-Aleut	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Haida	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Huavean	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Iroquoian	1	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Jicaquean	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kalapuyan	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Keresan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kiowa-Tanoan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lencan	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Maiduan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mayan	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3
Misumalpan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Miwok-Costanoan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mixe-Zoque	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3
Muskogean	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Otomanguean	0	1	1	1	1	1	1	2	2	3	3	4	4	5	6	6	6	7	7						
Palaihnihan	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pomoan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sahaptian	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Salishan	0	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	4	5						
Shastan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Siouan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
Tarascan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tequistlatecan	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Totonacan	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tsimshian	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Uto-Aztecan	1	1	1	1	1	1	1	1	2	2	3	4	4	4	5	5	6	6	7	7					
Wakashan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Wintuan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Xincan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yokutsan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yuki-Wappo	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
South America	**12**	**16**	**27**	**37**	**44**	**45**	**46**	**49**	**52**	**56**	**61**	**65**	**69**	**72**	**78**	**81**	**87**	**91**							
Araucanian	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1							

(continued)

Table A-2: (*continued*)

	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900
Arawakan	1	1	1	1	1	2	2	3	4	5	6	7	8	9	10	10	12	13
Arawan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Aymara	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Barbacoan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Boran	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bororoan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Cahuapanan	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Cariban	1	1	1	1	1	1	1	2	2	3	3	4	4	5	6	6	7	7
Chapacuran	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
Charruan	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Chibchan	0	1	1	1	1	1	1	1	1	1	2	2	2	2	3	3	3	3
Chiquitano	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Chocoan	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3
Chonan	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Guahibo	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Guaicuruan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hibito-Cholon	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Huarpean	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Huitotoan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Jivaroan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kakua-Nukak	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kamakanan	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kariri	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Katukinan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kawesqar	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lengua-Mascoy	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Matacoan	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Nadahup	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Nambiquaran	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Nuclear Macro-Je	0	1	1	1	1	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Panoan	1	1	1	1	1	1	1	1	2	2	3	3	3	4	4	4	4	4	5	5	5
Peba-Yagua	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Puri	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Quechuan	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	4	4	4	4	5
Saliban	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tacanan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ticuna-Yuri	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tucanoan	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3
Tupian	1	1	1	1	1	1	2	3	3	4	5	6	7	7	8	9	10	11			
Uru-Chipaya	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Yanomam	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Zamucoan	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Zaparoan	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Language isolate	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Unclassified	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

References

Bakker, Dik. 2011. Language sampling. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 100–127. Oxford: Oxford University Press.

Bell, Alan. 1978. Language samples. In Joseph H. Greenberg (ed.), *Universals of human language*, Vol. 1: *Method & theory*, 123–156. Stanford: Stanford University Press.

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11. 239–251

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

Bickel, Balthasar & Johanna Nichols. 2013. The Autotyp genealogy and geography database 2013 release. http://www.autotyp.uzh.ch/available.html

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.

Campbell, Lyle. 1997. *American Indian languages: The historical linguistics of Native America*. Oxford: Oxford University Press.

Croft, William. 2003. *Typology and universals*. 2nd edn. Cambridge: Cambridge University Press.

Cysouw, Michael. 2011. Understanding transition probabilities. *Linguistic Typology* 15. 415–431.

Dahl, Östen. 2008. An exercise in *a posteriori* language sampling. *Language Typology and Universals* 61. 208–220.

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

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

Dryer, Matthew S. 2000. Counting genera vs. counting languages. *Linguistic Typology* 4. 334–350.

Dryer, Matthew S. 2005a. Genealogical language list. In Haspelmath et al. (eds.) 2005, 584–644. Updates to the classification available in the online version of 2008 at http://blog.wals.info/errata-in-the-printed-edition-of-2005/

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

Dryer, Matthew S. 2013. Genealogical language list. In Dryer & Haspelmath (eds.) 2013. http://wals.info/languoid/genealogy

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max-Planck-Institut für evolutionäre Anthropologie. http://wals.info/

Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue: Languages of the world*. 15th edn. Dallas: SIL International. http://archive.ethnologue.com/15/web.asp

Grimes, Barbara F. (ed.). 1996. *Ethnologue: Languages of the world*. 13th edn. Dallas: Summer Institute of Linguistics.

Grimes, Joseph E. & Barbara F. Grimes. 1996. *Ethnologue: Language family index to the thirteenth edition of the Ethnologue*. Dallas: Summer Institute of Linguistics.

Hammarström, Harald. 2009. Sampling and genealogical coverage in *WALS*. *Linguistic Typology* 13. 105–119.

Hammarström, Harald & Mark Donohue. 2014. Some principles on the use of macro-areas in typological comparison. *Language Dynamics and Change* 4. 167–187.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2015. *Glottolog 2.4*. Leipzig: Max-Planck Institut für evolutionäre Anthropologie. http://glottolog.org

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

Henriksen, Carol & Johan van der Auwera. 1994. The Germanic languages. In Ekkehard König & Johan van der Auwera (eds.), *The Germanic languages*, 1–18. London: Routledge.

Himmelmann, Nikolaus P. 2000. Towards a typology of typologies. *Sprachtypologie und Universalienforschung* 53. 5–12.

Janhunen, Juha. 2009. Proto-Uralic – what, where, and when? In Jussi Ylikoski (ed.), *The quasquicentennial of the Finno-Ugrian Society* (Mémoires de la Société Finno-Ougrienne 258), 57–78. Helsinki: Finno-Ugrian Society.

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*, 615–750. Amsterdam: Benjamins.

Levinson, Stephen C., Simon J. Greenhill, Russell D. Gray & Michael Dunn. 2011. Universal typological dependencies should be detectable in the history of language families. *Linguistic Typology* 15. 509–534.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2015. *Ethnologue: Languages of the world*. 18th edn. Dallas: SIL International. http://www.ethnologue.com/

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4. 307–333.

Miestamo, Matti. 2003. *Clausal negation: A typological study*. Helsinki: Helsingin yliopisto doctoral dissertation.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter.

Miestamo, Matti. 2009. Implicational hierarchies and grammatical complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 80–97. Oxford: Oxford University Press.

Murdock, George Peter. 1968. World sampling provinces. *Ethnology* 7. 305–326.

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

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

Perkins, Revere D. 1992. *Deixis, grammar, and culture*. Amsterdam: Benjamins.

Perkins, Revere D. 2000. The view from hologeistic linguistics (Commentary on Maslova 2000). *Linguistic Typology* 4. 350–353.

Rankin, Robert L. 1993. On Siouan chronology. Paper presented at the Annual Meeting of the American Anthropological Association, Washington, DC.

Rijkhoff, Jan. 2009. On the (un)suitability of semantic categories. *Linguistic Typology* 13. 95–104.

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.

Ruhlen, Merritt. 1991. *A guide to the world's languages,* Vol. 1*: Classification, with a postscript on recent developments*. Stanford: Stanford University Press. Originally published in 1987 without postscript.

Stassen, Leon. 1997. *Intransitive predication*. Oxford: Oxford University Press.

Stolz, Thomas & Traude Gugeler. 2000. Comitative typology. *Sprachtypologie und Universalienforschung* 53. 53–61.
Tomlin, Russell S. 1986. *Basic word order: Functional principles*. London: Croom Helm.
Voegelin, Charles F. & Florence M. Voegelin. 1977. *Classification and index of the world's languages*. New York: Elsevier.
Wichman, Søren & David Kamholz. 2008. A stability metric for typological features. *Sprachtypologie und Universalienforschung* 61. 251–262.