

# Sampling and Genealogical Coverage in WALs

Harald Hammarström  
Dept. of Comp. Sci.  
Chalmers University  
412 96 Gothenburg  
SWEDEN  
`harald2@chalmers.se`

## Abstract

WALS was designed with the goal of providing a “systematic answer” to questions about the geographical distribution of language features. In order to achieve this goal, there must be an adequate sample of the world’s languages included in WALS. In this paper we investigate to what extent WALS fulfils its aim of maximizing the genealogical diversity of the samples of languages included. For this we look at the core 200-sample (included on almost all maps) as well as the 1370-sample for the feature OV/VO-word order (the sample with the largest number of languages). The genealogical diversity in these samples is compared against a database of “what could have been done”, i.e., a database of which language families have adequate descriptive resources for the task at hand. In the 200-sample, we find a highly significant overinclusion of Eurasian languages at the expense of South American and Papuan languages. In the 1370-sample, we find a highly significant overinclusion of North American languages at the expense of South American and Papuan languages. It follows that statistics based on these WALS samples cannot be used straightforwardly for sound inferences about the distribution of the features in question.

*Keywords:* Language Sampling, WALS, Language Classification, Linguistic Documentation, Basic Word Order

## Acknowledgements

I wish to thank Matthew Dryer for a number of objections raised to a near-final draft of this paper, some of which led to improvements in the explanation of the objective of this paper, and some of which merit a separate response from him. I wish to thank the following libraries for granting access and services: Centralbiblioteket (Gothenburg), Institutionen för orientaliska och afrikanska språk (Gothenburg), Etnografiska Muséet (Göteborg), LAI (Göteborg), Carolina Rediviva (Uppsala), NAI (Uppsala), Karin Boye (Uppsala), KB (Stockholm), SUB (Stockholm), LAI (Stockholm), Universiteitsbibliotheek (Leiden), KITLV (Leiden), Universiteitsbibliotheek (Amsterdam), Institute for Asian and African Studies (Helsinki), MPIEVA (Leipzig), Universitätsbibliothek (Leipzig), Butler/Columbia University (New York City), IfA (Cologne), BNF (Paris), SOAS (London), ILPGA (Paris), ZAS (Zürich). I am also indebted to (in no particular order) Hein van der Voort, Lincoln Almir Amarante Ribeiro, Eduardo Rivail Ribeiro, Michael Cysouw, Nathan Hill, Jesús Mario Girón, Karsten Legère, Helene Fatima Idris, Bernard Comrie, Lionel M. Bender, John Kalespi, Hilário de Sousa, Frank Seidel, Tom Güldemann, Lourens de Vries, Ian Tupper, Johanna Fenton, Randy Lebold, Willem Adelaar, Lyle Campbell, Norbert Cyffer, Maarten Mous, Thilo Schadeberg, Raoul Zamponi, Paul Whitehouse, Swintha Danielsen, Lauren Campbell, Dmitry Idiatov, Nick Evans, Matthew Dryer, Mark Donohue and Peter Bakker for help with access to data. The bibliographies by Alain Fabre (for South America) and Jouni Filip Maho (for Africa) have been very helpful in bibliographical searching leading up to this study.

## 1 Introduction

WALS<sup>1</sup> was designed with the goal of providing a “systematic answer” to questions about the geographical distribution of language features (Comrie et al. 2005a:1). As stressed in the introduction (Comrie et al. 2005a:1,4), in order to achieve this goal, there must be an adequate sample of the world’s languages included in WALS.

In this paper we will investigate to what extent WALS fulfils its goals

---

<sup>1</sup>For this paper, we used a hard copy with the accompanying CD-ROM purchased in August 2005 (Comrie et al. 2005b). If subsequent paper or web editions contain updates, they are not taken into account here.

and its claims, especially as it pertains to the desideratum of maximizing the genealogical diversity of the samples of languages included. For this we chose to investigate the core 200-sample and the 1370-sample for the feature OV/VO-word order. We chose the core 200-sample because this set was specially designed with the guiding principle to maximize genealogical diversity and the sampled languages are included on almost all maps. We chose the 1370-sample of the OV/VO-word order feature because it had the largest number of languages<sup>2</sup> included. The genealogical diversity in these samples will be compared against a database of “what could have been done”, i.e., a database of which language families have adequate descriptive resources for the task at hand. In other words, we contrast the “breadth” and “depth” of the genealogical diversity of WALS with the state-of-the-art possibility.

## 2 Preliminaries and Data Matters

### 2.1 WALS Languages

WALS contains ca 2560 languages<sup>3</sup> that appear on at least one map (Comrie et al. 2005a). Sign languages, pidgins and creoles have origins of a different kind, and fall outside the scope of this study (of genealogical coverage). There is a certain amount of language/dialect inconsistency, and a certain amount of coding inconsistency in WALS. However, in no case does this affect the genealogical classification of any WALS entry in this study, so such matters can be safely disregarded here.

### 2.2 Genealogical Classification

Without a full genealogical classification from the start, it is impossible to assess how well a certain sample covers the languages of the world. WALS provides a classification of the languages included in WALS, which is only a minority of languages in the world (Dryer 2005a). Thus, to assess the coverage it is necessary to use information from outside WALS.

WALS works with two different levels of genealogical classification; families “the highest level accepted by specialists” and genera “fairly obvious

---

<sup>2</sup>It is also the feature in WALS with the most genealogical diversity by any measure.

<sup>3</sup>The exact number is given as 2559 twice on page 3, but 2560 on page 4 and 584. From the data tables, 2560 appears to be the correct number of entries.

without systematic comparative analysis, and which even the most conservative ‘splitter’ would accept” (Dryer 2005a:584). We have chosen not to work with exactly these concepts, both for theoretical and practical reasons. In short, the family concept as of WALS is theoretically weak because it is not clear who is a specialist, that specialists agree, that all specialists know all the data, that specialists across areas have the same traditions on what to “accept”, and so on. In practice, there is no tangible evidence from specialists adduced to support the actual WALS-families as given – that is, there is no declaration of who the specialists are and where they argue their expert opinion, for each language family. Indeed, the outcome list honestly described as an “educated guess”. However, the outcome is quite different from, e.g., our educated guesses even about some undeniable trends in the opinion of the specialists.<sup>4</sup> Similarly, the genus concept lacks a threshold for “obviousness”<sup>5</sup> and there are practical problems here too – there is no evidence adduced to for the obviousness/non-obviousness of the various listed genera, and inconsistencies are easily spotted.<sup>6</sup>

For reasons just explained, it was infeasible for us to compile a list of genera or families in the WALS sense. Instead, we used the related concept of a “D-family” (for demonstrated family), defined as

- a **set of languages** (possibly a one-member set)
- with at least one **sufficiently attested** member language
- that has been **demonstrated in publication**
- to **stem from a common ancestor**

---

<sup>4</sup>To take just one example, Khoesan specialists agree that a ‘Khoisan’ family as listed in WALS, is **not** in evidence (Güldemann and Vossen 2000; Traill 1995; Westphal 1979). As the references in this paper show, there are very many more examples.

<sup>5</sup>Dryer (p.c. May 2008) admits that the “obviousness”-criterion may have to be adjusted to exclude the use of numerals as a criterion. The relatedness of most of the Indo-European branches can arguably be said to be obvious since several early amateurs independently saw the relatedness (using numerals as one of the arguments). If we want to call the various branches separate genera, then this fact must be dealt with – one way out is to rephrase “obvious” as “obvious considering everything except the numerals”.

<sup>6</sup>To take just one example here too, Central Solomons is listed as a genus, but these languages cannot even be shown to be related (Terrill 2006) while Germanic and Romance are listed as separate genera, despite the clear relatedness, as evidenced, for instance, in the numeral series.

- by **orthodox comparative methodology** (Campbell and Poser 2008).
- for which there are **no** convincing published attempts to demonstrate **a wider affiliation**

To support the actual choices, in each case, we give a reference to a publication pointing to the evidence necessary to establish the above, possibly adducing comments, in an appendix of supplementary online material to this paper.

While this concept is not free from theoretical or practical problems, we feel that it is preferable in terms of tangibility, as all choices are made more explicit. A fuller discussion is beyond the scope of this study. We feel that this practice is justified because, on the whole it does not really matter for the results what level of genealogical relatedness one counts, be it WALSGenera or D-families, as long as one is consistent across the world. As noted, we do not believe this property holds for specialists across different areas of the world, so it is not clear that WALSFamilies have this property.

### 2.3 Database of Descriptive Resources

According to WALSG (Comrie et al. 2005a:3), only 10-15 percent of the world’s languages are comprehensively described. However, no evidence, no distribution and no discussion surrounds this figure.

We have compiled a database of available descriptive resources for each D-family, similar to Hammarström (2007b) which is already becoming outdated. For each language family, one of the best descriptive resources for the best described language is listed, and categorized as “(Full) Grammar”, “Grammar sketch” and “Less than grammar sketch”. Ideally, the “(Full) Grammar” category would correspond to “sufficiently described for inclusion in the WALSG core-200 sample”, and “Grammar sketch” to “sufficiently described to decide the OV/VO feature”, but the matter is not so simple. Thus, we took care to check that our “(Full) Grammar” category included only languages for which the size of the description as a whole was similar to or larger than the least well-described languages actually included in the core-200 sample, and, for the question of the OV/VO feature, we took care to double check also exactly which grammar sketches give sufficient information to decide that feature, and which languages with poorer descriptive resources do in fact give sufficient information for OV/VO. The sources are listed explicitly in an appendix of supplementary online material to this paper.

There is a little discrepancy in that a few sources in our database (from 2008) became available too late for WALS. There are references in WALS from no later than 2004<sup>7</sup>, so we take anything from 2005 (inclusive) and on to have been 'too late' for WALS. When we evaluate the WALS samples, we take this into account, by "excusing" WALS in the posterior discussion if a source from after 2004 makes a difference. We keep good track of all such cases, and in no case do we fail to state them if they bear on the conclusions we draw.

Note also that some of the yet "poorly described" languages are extinct while others are not, which means that the status is subject to change in the future.

## 3 Evaluations of Coverage

### 3.1 The Core-200 Sample

The description of the construction of the core 200-sample is vague, but it is clear that it is not the result of a formal procedure. Rather it is an ad hoc procedure guided by the following principles: maximizing genealogical diversity and areal diversity, existence of a grammatical description, inclusion of major languages, inclusion of geographically disparate languages, hampered by availability of grammatical descriptions. Existence of a grammatical description must be understood as a mandatory criterion. The desiderata of maximizing genealogical and areal diversity are described with the word 'major', and the discourse indicates the remaining criteria to be minor (Comrie et al. 2005a:4-6).

---

<sup>7</sup>In this study, we used the break of 2004 to 2005 as the limit, for the reason just explained. Matthew Dryer (p.c. 2009) has since informed me that the core-200 sample was prepared in 1999, and therefore does not include languages for which a description appeared after 1999. This information was hardly deducible from WALS as it is not mentioned in the section about the core-200 sample, nor can it be inferred from the source lists for the core-200 languages, where, in several cases the best (or one of the best) source mentioned post-dates 1999, e.g., Ket (2000), Lepcha (2003), Shipibo-Conibo (2003), Apurinã (2000), Aymara (2001). The matter is not insignificant, because a lot of good descriptions for underdescribed families appeared in 2000-2004. Should the genealogical coverage of the core-200 sample be evaluated on the existence of grammatical descriptions as of the end of 1999 (rather than 2004, as in this paper), it seems that the Papuan and South American underinclusions would disappear, though we lack the database annotation needed to compute this exactly.

In spite of vagueness, we (and as we expect most WALS readers to do as well) find the following as the only consistent reading of the objective. A set of languages is selected, the members of which should be *maximally* genealogically and geographically diverse and, in addition, a sprinkle of further languages are added (by majorness and geographical disparateness) which do not increase to the genealogical and geographical diverseness of the whole set. The geographical diverseness maxim requires (at least) that large reasonably disjoint geographical regions are equally considered. The genealogical diverseness maxim requires that only languages from different families are selected. The goal of WALS, i.e., to provide a “systematic answer” to questions about the geographical distribution of language features, implies that no region should be overfocussed or underfocussed (that would hardly be systematic). The demands on systematicity and maximization leave no room for deviations. Thus, the objective of the core-200 sample must entail that languages included are from families evenly represented in large reasonably disjoint geographical regions. We now discuss whether this is indeed the case.

The 200-sample contains languages from 110 D-families. We recognize a total of 394 D-families in the world. However, only 212 of them contain a language for which there is a “(Full) Grammar”, i.e., is described comprehensively enough to be included in WALS on most features. In other words, WALS could be expanded to include most features for no less than  $212 - 110 = 112$  further language families.

However, it is not necessary to cover *all* possible families to achieve the goals of WALS, i.e., to provide a “systematic answer” to questions about the geographical distribution of language features, **as long as the families included are evenly sampled**. We will now go on to discuss whether this is true for the WALS core 200-sample.

Table 1 shows the continental break-up of the language family coverage of the WALS core 200-sample. Again, the figures refer to sufficiently well-described D-families versus D-families included in the WALS-200 sample. Well-described D-families that are not included in the WALS-200 sample are shown in bold.

As can be seen from Table 1, the sampling is not even across continents, ranging from 33.3% (Papua 12/36) to 84.6% (Eurasia 22/26) of the total number of families. If the inclusion of families were even across continents then all would have a coverage of around 51.9% (110/212). Are the differences we see statistically significant? We estimate the *p*-values by simulation as follows: 1. Generate 1000 random 110-member subsets  $S_i$  of the 212 well-

Africa 13/18 <b>72.2%</b>	Australia 11/21 <b>52.3%</b>	Eurasia 22/26 <b>84.6%</b>	North America 26/49 <b>53.0%</b>	Papua 12/36 <b>33.3%</b>	South America 26/62 <b>41.9%</b>
Afro-Asiatic	Bunaban	Abkhaz-Adyge	Algic	Angan	Araucanian
Atlantic-Congo	Gunwinyguan	Ainu	Caddoan	Austronesian	Arawa
Central Sudanic	Iwaidjan Proper	Austroasiatic	Cochimi-Yuman	Binanderean	Arawak
East Sudanic	Mangarrayi-Maran	Basque	Coosan	Border	Aymara
Furan	Minkin-Tangkic	Burushaski	Eskimo-Aleut	Lavukaleve	Barbacoan
Ju	Mirndi	Chukotko-Kamchatkan	Eyak-Athapaskan-Tlingit	Lower Sepik-Ramu	Bora-Huitoto
Kadugli-Krongo	Pama-Nyungan	Dravidian	Haida	Marind	Carib
Khoe-Kwadi	Tiwi	Indo-European	Iroquoian	Maybrat	Cayuvava
Kunama	Western Daly	Japanese	Karuk	Sentani	Chapacura
Maban	Worrorran	Kartvelian	Keresan	Sepik	Chibchan
Mande	Yangmanic	Korean	Kiowa-Tanoan	Toricelli	Chocoran
Saharan	<b>Anindilyakwa</b>	Miao-Yao	Kutenai	Trans New Guinea	Chonan
Songhay	<b>Eastern Daly</b>	Mongolian	Mayan	<b>Abun</b>	Guaicuruan
<b>Dogon</b>	<b>Gaagudju</b>	Nakh-Dagestanian	Miwok-Costanoan	<b>Awin-Pa</b>	Je
Ijoid	Garrwan	Nivkh	Mixe-Zoque	<b>Bilua</b>	Kawesqar
<b>Kuliak</b>	<b>Jarrakan</b>	Sino-Tibetan	Muskogean	<b>Cenderawasih Bay</b>	Matacoan
<b>Laal</b>	<b>Limilngan</b>	Tai-Kadai	Otomanguean	<b>East Bird's Head</b>	Mura-Piraha
<b>Omotic</b>	<b>Maningrida</b>	Tungusic	Pomoan	<b>Fasu</b>	Panoan
	<b>Nyulnyulan</b>	Turkic	Sahaptian	<b>Goilalan</b>	Peba-Yagua
	<b>Southern Daly</b>	Uralic	Salishan	<b>Hatam</b>	Quechuan
	<b>Wagiman</b>	Yeniseian	Siouan	<b>Inanwatan</b>	Tacanan
		Yukaghir	Tsimshian	<b>Kiwaian</b>	Trumai
		<b>Elamite</b>	Tunica	<b>Koarian</b>	Tucanoan
		<b>Hurro-Urartian</b>	Uto-Aztecan	<b>Kuot</b>	Tupi
		<b>Kusunda</b>	Wakshan	<b>Lower Mamberamo</b>	Warao
		<b>Sumerian</b>	Yuchi	<b>Nimboran</b>	Yanomam
			<b>Chimakuan</b>	<b>North Bougainville</b>	<b>Andoque</b>
			<b>Chimariko</b>	<b>North Halmahera</b>	<b>Bororo</b>
			<b>Chinook</b>	<b>Senagi</b>	<b>Cahuapanan</b>
			<b>Chumashan</b>	<b>Skó</b>	<b>Chiquitano</b>
			<b>Klamath-Modoc</b>	<b>South Bougainville</b>	<b>Fulnio</b>
			<b>Maiduan</b>	<b>Taulil-Butam</b>	<b>Guahibo</b>
			<b>Misumalpan</b>	<b>Teberan</b>	<b>Guato</b>
			<b>Molala</b>	<b>West Timor-Alor-Pantar</b>	<b>Harakmbut</b>
			<b>Salinan</b>	<b>Yale</b>	<b>Hibito-Cholon</b>
			<b>Seri</b>	<b>Yeli Dnye</b>	<b>Huarpean</b>
			<b>Shasta</b>		<b>Iranxe</b>
			<b>Siuslaw</b>		<b>Jabuti</b>
			<b>Takelma</b>		<b>Jivaro</b>
			<b>Tarascan</b>		<b>Kanoe</b>
			<b>Tequistlatecan</b>		<b>Karaja</b>
			<b>Timucua</b>		<b>Kwaza</b>
			<b>Tonkawa</b>		<b>Lengua-Mascoy</b>
			<b>Totonacan</b>		<b>Lule</b>
			<b>Wappo</b>		<b>Mochica</b>
			<b>Washo</b>		<b>Moseten-Chimane</b>
			<b>Wintuan</b>		<b>Movima</b>
			<b>Yokutsan</b>		<b>Nadahup</b>
			<b>Zuni</b>		<b>Nambiquaran</b>
					<b>Ofaie</b>
					<b>Paez</b>
					<b>Puelche</b>
					<b>Puinave</b>
					<b>Saliban</b>
					<b>Taushiro</b>
					<b>Ticuna</b>
					<b>Urarina</b>
					<b>Uru-Chipaya</b>
					<b>Waorani</b>
					<b>Yamana</b>
					<b>Yurakare</b>
					<b>Zamucoan</b>

Table 1: Continental break-up of the language family coverage of the WALS core 200-sample. The figures refer to sufficiently well-described D-families versus D-families included in the WALS-200 sample. Well-described D-families that are not included in the WALS-200 sample are shown in bold.

Continent	W200	Question	Test	Outcome	$p$ -value
Eurasia	22/26	Overinclusion	$ \{i S_i[\text{Eurasia}] \geq 22\} $	0	$p < 0.001$
Africa	13/18	Overinclusion	$ \{i S_i[\text{Africa}] \geq 13\} $	13	$p \approx 0.063$
South America	26/62	Underinclusion	$ \{i S_i[\text{South America}] \leq 26\} $	37	$p \approx 0.037$
Papua	12/36	Underinclusion	$ \{i S_i[\text{Papua}] \leq 12\} $	13	$p \approx 0.013$

Table 2: Tests for statistical significance of over/under-inclusion in the WALS-200 sample. All  $i$ :s range up to 1000.  $S_i[C] = |\{x|x \in S_i \text{ such that } x \text{ is from continent } C\}|$ .

described families; 2. For each family, ask how many of the subsets have more/less than the number found in the WALS-200 sample. For instance, if, say, 100 of those 1000 random subsets contain more Eurasian languages than in the WALS-200 sample, then there is no statistically significant over-sampling from Eurasia in the WALS-200 sample, because higher inclusion of Eurasian families are included too often just by random. The results are shown in Table 2.

In plain words, we find that overinclusion of Eurasian D-families is highly significant in the WALS-200 sample, and that this happens at the expense of South American and Papuan D-families (but the significance levels in of these underinclusions are much lower). Overinclusion of African D-families is not significant with conventional levels of significance. The underinclusion of Papuan languages is somehow excusable because of a number of borderline choices. Firstly, based mostly on pronominal morphemes, Abun, Hatam and North Halmahera are widely held to be genetically related, in spite of the lacking lexical correspondences (Klamer et al. 2008). Secondly, extensive descriptions of Inanwatan and Sko member languages did not appear until too late for WALS, and for Awin-Pa languages there were text collections out (Stewart 1987), but lengthy written-up grammar papers were not easily accessible until recently (2008), when Routamaa (1994) was posted online. Thirdly, a few more Papuan D-families which are here listed as being sufficiently well-described, are actually debatable (though still comparable in descriptive status to the least well-described Papuan languages in the WALS-200 sample). If these borderline cases are turned in favour of WALS, the Papuan coverage is well within non-significance limits.

The vast overinclusion of Eurasian D-families is disturbing, especially since the pitfalls of Eurasian oversampling is precisely what is highlighted in the WALS sampling section (Comrie et al. 2005a:3)! Note that this fact

has nothing to do with the desideratum of including extra “major” languages (Comrie et al. 2005a:3), because, e.g., the Eurasian families Abkhaz-Adyge, Ainu, Burushaski, Basque, Chukotko-Kamchatkan, Miao-Yao, Nivkh, Tungusic, Yeniseic and Yukaghir contain no major language in terms of speaker numbers (Gordon 2005). Furthermore, we may look at the Eurasian families which were not included; they are the ancient families, all of them long extinct<sup>8</sup> – ancient families were consciously excluded from WALS on other considerations – and the isolate Kusunda, for which a full-ish description only became available too late for WALS (Watters 2005). Thus, considering that the WALS core-200 sampling was designed never to catch these families, the Eurasian bias in the selection set is even stronger than the presented figures (every available Eurasian family was caught!).

### 3.2 The 1370-OV/VO Sample

As mentioned already, the OV/VO feature is the feature in WALS with the largest number of languages included (and also the feature which includes languages for the largest number of D-families). It is not clear how the data points/languages were selected, but it may be guessed that it is some kind of convenience sample (Dryer 2005b).

Again, it may be interesting to see to what the languages included are evenly sampled. The 1370-sample contains languages from 244 D-families. We recognize a total of 394 D-families in the world. However, only 339 of them contain a language for which there is a publication with information to decide the OV/VO feature. In other words, WALS could be expanded to include the OV/VO feature for no less than  $339 - 244 = 95$  further language families.

However, it is not necessary to cover *all* possible families to achieve the goals of WALS, i.e., to provide a “systematic answer” to questions about the geographical distribution of language features, **as long as the families included are evenly sampled**. We will now go on to discuss whether this is true for the WALS-OV/VO 1370-sample.

Tables 3-4 show the continental break-up of the language family coverage of the WALS-OV/VO 1370-sample. The figures refer to the number of D-families for which there is a publication with information to decide the

---

<sup>8</sup>One may also perhaps exclude Hurro-Urartian and Elamite, as we do with Etruscan, on the grounds that they are not sufficiently well-known.

OV/VO feature, versus D-families included in the WALS-OV/VO 1370 sample. The D-families with a description that are not included in the WALS-OV/VO 1370 sample are shown in bold.

As can be seen from Tables 3–4, the sampling is not even across continents, ranging from 59.7% (Papua 55/92) to 86.1% (North America 56/65). If the inclusion of families were even across continents then all would have a coverage of around 72.0% (244/339). Are the differences we see statistically significant? Again, we estimate the  $p$ -values by simulation as follows: 1. Generate 1000 random 244-member subsets  $S_i$  of the 339 families for which sufficient information exists; 2. For each family, ask how many of the subsets have more/less than the number found in the WALS-OV/VO 1370 sample. The results are shown in Table 5.

In plain words, we find that overinclusion of North American D-families is highly significant in the WALS-OV/VO 1370 sample, and that this happens at the expense of South American and Papuan D-families (but the significance levels in of these underinclusions are lower). Overinclusion of Australian D-families is not significant with conventional levels of significance (Eurasian even less so).

Again, in fact, the underinclusion of Papuan D-families is excusable for the following reasons. The data for the Kwomtari, Baibai, Guriaso and Yuat-Maramba languages is difficult to access (or too recent). The published basis available for assignment in the Pahoturi, Bayono-Awbono, Elseng, Abinomn and Lepki is just a plain statement of the basic word order, without any explicit source or examples – a prudent researcher would perhaps require at least a sentence example to be convinced. For Amtom-Musan only unanalysed text material is available (Krieg 1992) and in the case of Waia and Inanwatan, material is too recent for WALS. If we remove these 12 families from the list of sufficiently described families, 327 families remain in total, 80 (rather than 92) being Papuan. The inclusion proportion for Papuan languages rises from 59.7% (55/92) to 68.7% (55/80), and most importantly, underexclusion for Papuan families is no longer statistically significant ( $p \approx 0.118$ ).

There are no similar considerations which would explain the dearth of South American families. It is clear that the neglect of South American languages lamented two decades ago by Derbyshire and Pullum (1986) is not quite over yet.

Africa 28/35 <b>80.0%</b>	Australia 22/26 <b>84.6%</b>	Eurasia 25/31 <b>80.6%</b>	North America 56/65 <b>86.1%</b>	Papua 55/92 <b>59.7%</b>	South America 58/90 <b>64.4%</b>
Afro-Asiatic	Anindilyakwa	Abkhaz-Adyge	Algic	Abun	Araucanian
Atlantic-Congo	Anson Bay	Ainu	Alea	Anem	Arawa
Berta	Bunaban	Austroasiatic	Atakapan	Angan	Arawak
Central Sudanic	Eastern Daly	Basque	Caddoan	Austronesian	Atacame
Dogon	Garrwan	Burushaski	Chimakuan	Awin-Pa	Atacameño
East Sudanic	Gunwinyguan	Chukotko-Kamchatkan	Chimariko	Bilua	Aymara
Furan	Iwaidjan Proper	Dravidian	Chinook	Binanderean	Barbacoan
Gumuz	Larrakiyan	Great Andamanese	Chitimacha	Border	Bora-Huitoto
Heiban	Limilngan	Indo-European	Chumashan	Bosavi	Bororo
Ijoid	Mangarrayi-Maran	Japanese	Coahuilteco	Bulaka River	Cahuapanan
Ju	Maningrida	Jarawa-Onge	Cochimi-Yuman	Cenderawasih Bay	Candoshi-Shapra
Kadugli-Krongo	Minkin-Tangkic	Kartvelian	Coosan	Duna-Bogaya	Candoshi-Shapra
Katla-Tima	Mirndi	Korean	Eskimo-Aleut	East Bird's Head	Cayuvava
Khoe-Kwadi	Northern Daly	Kusunda	Eyak-Athapaskan-Tlingit	East Kutubu	Chapacura
Koman	Nyulnyulan	Miao-Yao	Haida	Eastern Trans-Fly	Chibchan
Kuliak	Pama-Nyungan	Mongolian	Huavean	Fasu	Chichitano
Kunama	Southern Daly	Nakh-Dagestanian	Iroquoian	Goilalan	Chococoan
Laal	Tiwi	Nivkh	Karakawa	Hatam	Chonan
Maban	Wagiman	Sino-Tibetan	Karuk	Kaki Ae	Fulnio
Mande	Western Daly	Tai-Kadai	Keresan	Karkar	Guahibo
Narrow Talodi	Worrorrnan	Tungusic	Kiowa-Tanoan	Kayagar	Guaicuruan
Omotic	Yangmanic	Turkic	Klamath-Modoc	Kiwaian	Guato
Rashad	<b>Gaagudju</b>	Uralic	Kutenai	Koarian	Hibito-Cholon
Saharan	<b>Giimbiyu</b>	Yeniseian	Maiduan	Kolopom	Iranxe
Shabo	<b>Jarrakan</b>	Yukaghir	Mayan	Kuot	Itonama
Songhay	<b>Umbugarla</b>	<b>Elamite</b>	Misumalpan	Kwerba	Jabuti
Tegem		<b>Etruscan</b>	Miwok-Costanoan	Lakes Plain	Je
Tuu		<b>Hattic</b>	Mixe-Zoque	Lavukaleve	Jirajaran
<b>Bangi Me</b>		<b>Hurro-Urartian</b>	Muskogean	Lower Mamberamo	Jivaro
<b>Hadza</b>		<b>Nihali</b>	Natchez	Lower Sepik-Ramu	Karaja
<b>Hoa</b>		<b>Sumerian</b>	Otomanguean	Mairasi	Kariri
<b>Mao</b>			Palaihnihan	Marind	Kawesqar
<b>Meroitic</b>			Pomoan	Maybrat	Matacoan
<b>Ongota</b>			Sahaptian	Mombum	Moseten-Chimane
<b>Sandawe</b>			Salinan	Moraori	Mura-Piraha
			Salishan	Morehead UM Rivers	Nadahup
			Seri	Mpur	Nambiquaran
			Shasta	North Halmahera	Paez
			Siouan	Oksapmin	Panoan
			Siuslaw	Pawaia	Peba-Yagua
			Takelma	Senagi	Quechuan
			Tarascan	Sentani	Rikbaktsa
			Tequistlatecan	Sepik	Saliban
			Timucua	Sko	Tacanan
			Tonkawa	South Bougainville	Taushiro
			Totonacan	Suki-Gogodala	Ticuna
			Tsimshian	Sulka	Timote-Cuica
			Tunica	Tor-Orya	Trumai
			Uto-Aztecan	Torricelli	Tucanoan
			Wakashan	Trans New Guinea	Tupi
			Wappo	Turama-Kikori	Urarina
			Washo	West Bomberai	Uru-Chipaya
			Wintuan	West Timor-Alor-Pantar	Waorani
			Yokutsan	Yale	Warao
			Yuchi	Yeli Dnye	Yamana
			Zuni	<b>Abinomn</b>	Yanomam
			<b>Comecrudan</b>	<b>Amto-Musan</b>	Zamucoan
			<b>Guaicurian</b>	<b>Arafundi</b>	Zaparoan
			<b>Jicaquean</b>	<b>Ata</b>	<b>Aikana</b>
			<b>Kalapuyan</b>	<b>Baibai</b>	<b>Aimore</b>
			<b>Lencan</b>	<b>Baining</b>	<b>Andoque</b>
			<b>Molala</b>	<b>Bayono-Awbono</b>	<b>Awake</b>
			<b>Xincan</b>	<b>Burmeso</b>	<b>Betoi-Jirara</b>
			<b>Yana</b>	<b>Dem</b>	<b>Charrua</b>
			<b>Yuki</b>	<b>East Timor</b>	<b>Chono</b>

Table 3: Continued on the next page.

Africa contd.	Australia contd.	Eurasia contd.	North America contd.	Papua contd.	South America contd.
				<b>Elemán</b>	<b>Cofán</b>
				<b>Elseng</b>	<b>Harakmbut</b>
				<b>Guriaso</b>	<b>Huarpean</b>
				<b>Inanwatan</b>	<b>Jodi</b>
				<b>Kaure-Kapori</b>	<b>Kakua-Nukak</b>
				<b>Kol</b>	<b>Kamsa</b>
				<b>Kolana-Tanglapui</b>	<b>Kanoe</b>
				<b>Konda-Yahadian</b>	<b>Katukina</b>
				<b>Kwomtari</b>	<b>Kwaza</b>
				<b>Left May</b>	<b>Leko</b>
				<b>Lepki</b>	<b>Lengua-Mascoy</b>
				<b>Masep</b>	<b>Lule</b>
				<b>Nimboran</b>	<b>Maku</b>
				<b>North Bougainville</b>	<b>Maxakali</b>
				<b>Pahoturi</b>	<b>Mochica</b>
				<b>Piawi</b>	<b>Movima</b>
				<b>Savosavo</b>	<b>Muniche</b>
				<b>South Bird's Head Proper</b>	<b>Ofaie</b>
				<b>Taiap</b>	<b>Puelche</b>
				<b>Taulil-Butam</b>	<b>Puinave</b>
				<b>Teberan</b>	<b>Puquina</b>
				<b>Touo</b>	<b>Tinigua</b>
				<b>Uhunduni</b>	<b>Vilela</b>
				<b>Waia</b>	<b>Yaruro</b>
				<b>West Bird's Head</b>	<b>Yurakare</b>
				<b>Yawa</b>	
				<b>Yuat-Maramba</b>	

Table 4: The continental break-up of the language family coverage of the WALS-OV/VO 1370-sample. The figures refer to the number of D-families for which there is a publication with information to decide the OV/VO feature, versus D-families included in the WALS-OV/VO 1370 sample. The D-families with a description that are not included in the WALS-OV/VO 1370 sample are shown in bold.

Continent	W1370	Question	Test	Outcome	$p$ -value
North America	56/65	Overinclusion	$ \{i S_i[\text{North America}] \geq 56\} $	1	$p \approx 0.001$
Australia	22/26	Overinclusion	$ \{i S_i[\text{Eurasia}] \geq 22\} $	90	$p \approx 0.090$
South America	58/90	Underinclusion	$ \{i S_i[\text{South America}] \leq 58\} $	41	$p \approx 0.041$
Papua	55/92	Underinclusion	$ \{i S_i[\text{Papua}] \leq 55\} $	3	$p \approx 0.003$

Table 5: Tests for statistical significance of over/under-inclusion in the WALS-OV/OV 1370 sample. All  $i$ :s range up to 1000.  $S_i[C] = |\{x|x \in S_i \text{ such that } x \text{ is from continent } C\}|$ .

## 4 Discussion

WALS is aimed at giving researchers a tool to investigate frequencies and correlations of language features (Comrie et al. 2005b:1). For this to be meaningful, the WALS data, at least if further refined, should be sufficient for approaching some level of inferential validity, without recourse to collecting further data.

We have shown that there are significant gaps in the breadth and depth coverage of the WALS-data. Whether these gaps alter conclusions drawn in studies based on the WALS-data depends on the specifics of the individual studies. What we indicate here is that the WALS-data cannot be used blindly to draw statistically valid inferences about the state-of-the-art knowledge of the world’s languages.

There is a further point not to be forgotten as to the WALS-data and statistical inferences, which has traditionally been overlooked in typology (but cf. Hammarström 2007a and Janssen et al. 2006). Statistically valid conclusions about a population can only be drawn if data is sampled *at random* from the population. Whatever the method(s) to used to include languages in WALS, it was not random, in the sense of picking from a plain list of languages, from lists of languages stratified into genera, or from a list of languages stratified into families. It is discernable that WALS-data points were included based on convenient availability of data, and this may or may not turn out to be functionally equivalent (in terms of feature distributions) to a random selection. To test whether this is the case is beyond the scope of thus study, but we may nevertheless speculate on how a skewing may come about; languages which are deep in the lowlands of New Guinea and deep in the Amazon forest may be less influenced by SVO contact languages, Malay/Indonesian and Spanish/Portuguese respectively, than their more accessible more documented counterparts. Now, of course, typologists must

use data from documented languages rather than undocumented ones – we are certainly not attempting to imply any fault here – what we do wish to remind of, is that blind statistical inferences onto the world’s languages based on such data, are not necessarily sound.

## 5 Conclusion

We have shown that even in the cases of the broadest and deepest data in WALS, they are not a perfect mirror of the genealogical diversity of the languages of the world as of today’s knowledge. In the 200-sample, there is a strong bias favouring Eurasian families, and in the 1370-sample there is a strong bias favouring North American families. In both cases, there is also a weak but statistically significant bias disfavouring Papuan and South American families. In both cases, the WALS may be “excused” for the underinclusion of the Papuan families, but in neither case is the underexclusion of South American families or the overinclusion of Eurasian/North American families justifiable. The study depends a little on classification of data size and availability, the demarcation of which is not watertight, and as more data becomes available in the future, the situation will change further. One may also discuss the level and detail of genealogical classification used here, but even so, the conclusions above are likely to remain nevertheless. Caution is due when using the WALS-data to draw statistical inferences.

## References

- Campbell, L. and Poser, W. J. (2008). *Language Classification: History and Method*. Cambridge University Press.
- Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M. (2005a). Introduction. In Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M., editors, *World Atlas of Language Structures*, pages 1–8. Oxford University Press.
- Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M., editors (2005b). *World Atlas of Language Structures*. Oxford University Press. xv+695pp [A3 size].
- Derbyshire, D. C. and Pullum, G. K. (1986). Introduction. In Derbyshire,

- D. C. and Pullum, G. K., editors, *Handbook of Amazonian Languages*, volume I, pages 1–30. Mouton de Gruyter.
- Dryer, M. S. (2005a). Genealogical language list. In Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M., editors, *World Atlas of Language Structures*, pages 584–644. Oxford University Press.
- Dryer, M. S. (2005b). Order of object and verb. In Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M., editors, *World Atlas of Language Structures*, pages 338–341. Oxford University Press.
- Gordon, Jr., R. G., editor (2005). *Ethnologue: Languages of the World*. Dallas: SIL International, 15 edition.
- Güldemann, T. and Vossen, R. (2000). Khoisan. In Heine, B. and Nurse, D., editors, *African Languages: An Introduction*, pages 99–122. Cambridge University Press.
- Hammarström, H. (2007a). A genetically stratified language sample for basic word order typology. Paper Presented at The seventh International Conference of the Association for Linguistic Typology (ALT VII), CNRS, Paris, September 25-28, 2007.
- Hammarström, H. (2007b). *Handbook of Descriptive Language Knowledge: A Full-Scale Reference Guide for Typologists*, volume 22 of *LINCOM Handbooks in Linguistics*. München: Lincom.
- Janssen, D. P., Bickel, B., and Zúñiga, F. (2006). Randomization tests in language typology. *Linguistic Typology*, 10:419–420.
- Klamer, M., Reesink, G., and van Staden, M. (2008). East Nusantara as a linguistic area. In Muysken, P., editor, *From linguistic areas to areal linguistics*, volume 90 of *Studies in Language Companion Series*, pages 95–149. Amsterdam: John Benjamins.
- Krieg, L. (1992). *Tienesi [Siawi Genesis]*. Goroka, Papua New Guinea: New Tribes Mission.
- Routamaa, J. (1994). Kamula grammar essentials. Ms. Available at <http://www.sil.org/pacific/png/abstract.asp?id=50209> accessed 1 August 2008.

- Stewart, J. (1987). *God ya tyo kimina, God ya swagumin nin [New Testament in Aekyom]*. Port Moresby: Bible Society Papua New Guinea.
- Terrill, A. (2006). Central Solomon languages. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, volume 2, pages 279–281. Amsterdam: Elsevier, 2 edition.
- Traill, A. (1995). The Khoesan languages of South Africa. In Mesthrie, R., editor, *Language and social history: studies in South African sociolinguistics*, pages 1–18. Cape Town: David Philip.
- Watters, D. E. (2005). *Notes on Kusunda Grammar: A Language Isolate of Nepal*, volume 3 of *Himalayan Linguistics Archive*. National Foundation for the Development of Indigenous Nationalities.
- Westphal, E. O. J. (1979). Languages of Southern Africa. In *Perspectives on the Southern African past*, volume 2 of *Occasional papers / Centre for African studies, University of Cape Town*, pages 37–58. Rondebosch.