3

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 34–71
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4560

Systematic typological comparison as a tool for investigating language history

Ger Reesink

Max Planck Institute, Nijmegen

Michael Dunn

Max Planck Institute, Nijmegen

Similarities between languages can be due to 1) homoplasies because of a limited design space, 2) common ancestry, and 3) contact-induced convergence. Typological or structural features cannot prove genealogy, but they can provide historical signals that are due to common ancestry or contact (or both). Following a brief summary of results obtained from the comparison of 160 structural features from 121 languages (Reesink, Singer & Dunn 2009), we discuss some issues related to the relative dependencies of such features: logical entailment, chance resemblance, typological dependency, phylogeny and contact. This discussion focusses on the clustering of languages found in a small sample of 11 Austronesian and 8 Papuan languages of eastern Indonesia, an area known for its high degree of admixture.

1. Introduction. The practice of proposing families on the basis of typological comparison is one of the guilty secrets of historical linguistics. It is a basic principle of the historical linguistic tradition that genealogical relationships between languages can only be established by the comparative method, which detects sets of cognates on the basis of regular sound changes and shared irregularities, and thus allows the positing and reconstruction of a proto-language[1]. In spite of this, some early classifications of the more than 800 Papuan languages are based on just a handful of lexical correspondences, supplemented by observations of structural and typological similarities (Greenberg 1971;

[1] The original research conducted for this study was supported by funds from NWO (Netherlands Organization for Scientific Research) for the Program "Breaking the time barrier: Structural traces of the Sahul past" of Professor Pieter Muysken (360-70-210), Radboud University Nijmegen and Professor Stephen C. Levinson, Max Planck Institute for Psycholinguistics, Nijmegen. We thank two anonymous reviewers for comments on an earlier version and Angela Terrill for editorial assistance.

Wurm 1975, 1982). These proposals have been severely criticized (see Pawley 1998, 2005 for a summary), but the influence of typological data at the stage of genealogical hypothesis generation remains.

Typological features of languages are subject to the same evolutionary processes which create genealogical history in other aspects of samples of related languages. There is a tendency for more closely related languages to be more similar on the level of linguistic structure, just as they are more similar in terms of e.g. shared vocabulary. The evolutionary and statistical properties of lexical and sound change have been extensively examined: a great deal is known about what kinds of sound changes are likely, as there is too about what kinds of words tend to be lost, replaced, semantically or phonologically mutated, and so forth. Less is known about the evolutionary and statistical properties of typological/structural features, Even where lexical cognates cannot be identified because of phonological and semantic drift, there remains the possibility that other aspects of language retain traces of the historical relations between languages, whether due to genealogical descent or contact. Area specialists may be able to make generalizations about languages of one or another family on the basis of typological features even where comparative method reconstruction has not been carried out. Hypothesis generation on the basis of structural features of language relies intrinsically on statistical arguments.

As in biology, there are a number of different historical factors that lead languages to be similar: common ancestry, contact (hybridization), and chance convergence (homoplasy). The smaller the design space the higher the probability that convergence is the result of chance rather than geneaological or geographical factors. In biological evolution therefore, the more degrees of freedom in a given domain, the more powerful is the mutation and selection process, resulting in greater disparity and diversity of species. This suggests for linguistic evolution that the greater degree of freedom of lexical elements allows for a more exact measure of phylogenetic relationship on the basis of cognacy sets. Structural features have a much more limited design space, thus convergent evolution will cause homoplasies that need to be distinguished from historical signals, be they phylogenetic or due to hybridization. Large scale chance convergence is less likely, however, when a great number of features are compared, provided these have a measure of independence. See for a more extensive argumentation Dunn et al. (2008:715) where we answer the skepticism expressed by Harrison (2003). We come back to this point in the conclusion.

In this paper we examine the statistical properties of structural features of languages with an eye to their potential in illuminating historical relations. We use the languages of eastern Indonesia, previously identified as an interesting area including both diffusion and inheritance, as a case study. We identify various traits of these languages as present either through diffusion or genetic inheritance.

We adopt a systematic, probabilistic approach using computational models. There are a number of reasons for this, both practical and theoretical. Practically, computational models are able to process a multitude of traits for a great number of languages, while minimizing the apophenic effects of observer preconceptions , where 'apophenic' refers to the human tendency to see meaningful patterns or connections in random or meaningless data. Theoretically, computational models provide us with consistent and testable results, comparable over different hypotheses, and having useful statistical properties such as explicit likelihood scores. A further advantage of computational methods over the

Comparative Method is that the former approach allows hypothesis generation and testing in a way not possible with the Comparative Method. We show that while the Comparative Method illuminates genealogy, structural features can illuminate a long-term history of contact.

The use of structural data in phylogenetic inference has been applied in a few earlier studies which are summarized in section 2. In section 3 we discuss the number and nature of structural features that have been used in those studies. In particular, we pay attention to the issue of trait independency. Section 4 presents the results of a small-scale study, illustrating how structural features provide some clusterings in a set of genealogically diverse Austronesian and Papuan languages of eastern Indonesia. Here we attempt to distill which set of features contributes most strongly to the clusterings. The conclusion in section 5 summarizes discoveries and remaining issues of a standardized approach to typological comparison.

2. Previous studies employing structural features. The use of structural data in phylogenetic inference has been applied in an investigation into the relationships between twenty-two languages of the Oceanic subgroup of the Austronesian family and fifteen Papuan languages of Island Melanesia, reported in two publications (Dunn et al. 2005, Dunn et al. 2008). Although the Papuan languages of this sample had been claimed to form a genealogical group (the East-Papuan phylum, see Wurm 1975), this genealogical unity had been challenged by Ross (2001) and Dunn et al. (2002).

Dunn et al. (2005) used a maximum parsimony analysis of the distribution of 125 abstract structural features and found a reasonable congruence between the consensus tree and the traditional classification of the Oceanic languages in their sample, while the Papuan tree showed some geographic clustering, possibly reflecting ancient relationships (due to inheritance or diffusion through contact). For a critical debate on the merits of that study see Donohue and Musgrave (2007) and Dunn et al. (2007). Croft (2008:230) remarks, "although the result from Dunn et al. (2005) is surprising to a historical linguist, it may be that a cluster of typological traits will provide more precision in classification than will individual traits; also some typological traits are quite stable and therefore may be useful indicators of phylogeny."

Dunn et al. (2008) explained various computational methods in more detail, showed how they can be extended and refined and explored how a phylogenetic signal can be distinguished from possible contact. That study used a Bayesian algorithm to carry out a phylogenetic analysis on a set of 115 abstract phonological and grammatical features. While a certain degree of possible admixture of structural features was detectable between some Oceanic and some Papuan languages, the overal clustering of the languages distinguished the Papuan languages from the Oceanic languages, and the Papuan languages could be clustered into three (geographically, archaeologically) plausible subgroups. The clustering of the Papuan languages into three groups was shown not to be the result of degrees of contact with Oceanic languages, leaving as the most plausible hypothesis that the historical signal found on the basis of structural features is most likely due to a common ancestry, ancient contact between Papuan lineages, or both.

One of the questions raised by these studies (Dunn et al. 2008:737) was how the eastern Papuan languages of Island Melanesia would cluster if a much greater sample

of Papuan languages were investigated. In their critique on Dunn et al. (2005) Donohue and Musgrave (2007:11) "proposed that comparison with Austronesian languages should include representative Austronesian languages from beyond Island Melanesia, in order to obtain an idea of the degree of diversity of these features that can be expected in a family over a 10,000 year (in the Austronesian case, 6,000 year) time frame."

For a follow-up study designed to apply the structural method to a much larger sample of languages, the set of structural features was critically reviewed, revised and expanded. See below for a comparison of some revised questions and the Appendix for both questionnaires.

In the second study (Reesink et al. 2009) we compared a large sample of 121 languages from the Sahul region (i.e. New Guinea and Australia), made up of 55 Papuan, 17 Australian and 48 Austronesian languages, and one Andamanese language, using the revised and expanded set of 160 structural features. Since the linguistic situation of Sahul is complex, combining great time depth with long-term and intensive contact situations, we used a Bayesian algorithm originally developed to discover population structure on the basis of recombining genetic markers, i.e. a model of inheritance and admixture. The Structure algorithm (Pritchard et al. 2000) models evolutionary change and admixture and simultaneously determines both the most likely number of ancestral groups and the most likely contribution of each of these ancestral populations to each of the observed individuals (in this case, languages). The results of Reesink et al. (2009) study suggest 10 ancestral linguistic populations, some of which largely correspond to clearly defined or proposed phylogenetic groups (see figure 1), while others exhibit a high degree of hybridization. Where there are very different degrees of hierarchical relatedness the inferred populations may be nested within known genealogical groupings. The 10 ancestral populations inferred by the structure algorithm can be characterized as follows:

The **Austronesian family** is captured by three groups:

dark green	The Austronesian languages of Borneo and the Phillipines
pale blue	Oceanic languages of mainland New Guinea, New Britain, and Vanuatu
dark purple	All other Oceanic languages of the sample

The Tsou language of Taiwan is equally related to the *dark green* and *dark purple* groups

Other major families

dark blue	Trans-New-Guinea (note that this does not include some of the languages hypothesised to belong to the TNG periphery, such as the Alor-Pantar languages)
light green	Pama-Nyungan languages

Areal groupings

light orange	Non-Pama-Nyungan languages
dark orange	North coast Papuan
light purple	South coast Papuan

pink	East Papuan (plus Bukiyip and Yimas in the north of New Guinea)
red	West Papuan (the Alor-Pantar languages, plus some difficult to classify languages of Halmahera)

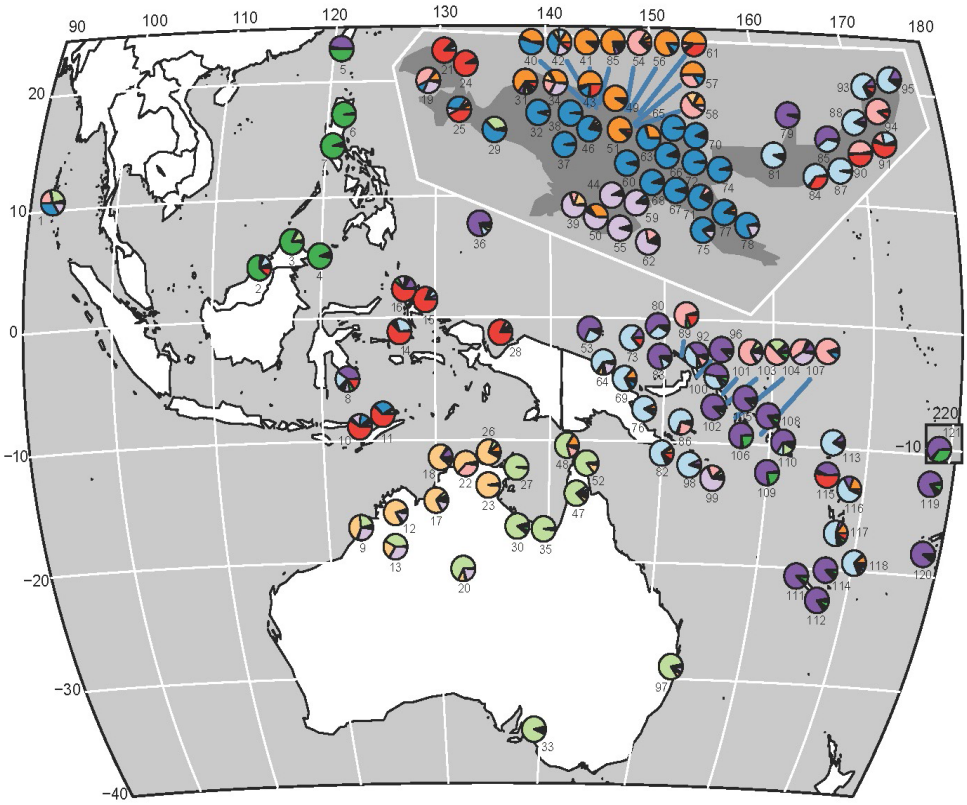

FIGURE 1. The geographic patterning of Structure results for 10 founding populations (Reesink et al. 2009). The pie charts indicate the proportional contribution of each of the founding populations to each language. Languages are identified by number:

Legend

(Fuller details of the interpretation of each population are given in Reesink et al. 2009: 4-7.)

1. Onge [oon]
2. Belait [beg]
3. Kimaragang [kqr]
4. Sama [ssb]
5. Tsou [tsu]
6. Ilocano [ilo]
42. Imonda [imn]
43. Isaka [ksi]
44. Arammba [stk]
45. Namia [nnm]
46. Telefol [tlf]
47. Kuuk Thayorre [thd]
83. Tungag [lcm]
84. Mangseng [mbh]
85. Nakanai [nak]
86. Kilivila [kij]
87. Mengen [mee]
88. Meramera [mxm]

7. Tagalog [tgl]
8. Muna [mnb]
9. Bardi [bcj]
10. Klon [kyo]
11. Abui [abz]
12. Ngarinyin [ung]
13. Gooniyandi [gni]
14. Taba [mky]
15. Tidore [tvo]
16. Tobelo [tlb]
17. Murrinhpatha [mwf]
18. Tiwi [tiw]
19. Inanwatan [szp]
20. Warlpiri [wbp]
21. Meyah [mej]
22. Mawng [mph]
23. Bininj Gun-wok [gup]
24. Hatam [had]
25. Mairasi [zrs]
26. Burarra [bvr]
27. Djambarrpuyngu [djr]
28. Biak [bhw]
29. Kamoro [kgq]
30. Garrwa [gbc]
31. Bauzi [bvz]
32. Nggem [nbq]
33. Ngarrinyeri [nay]
34. Orya [ury]
35. Kayardild [gyd]
36. Ulithian [uli]
37. Korowai [khe]
38. Una [mtg]
39. Marind [mrz]
40. Menggwa Dla [kbv]
41. Abau [aau]
48. Kala Lagaw Ya [mwp]
49. Mende [sim]
50. Gizrra [tof]
51. Yessan-Mayo [yss]
52. Uradhi [urf]
53. Wuvulu-Aua [wuv]
54. Bukiyip [ape]
55. Bine [bon]
56. Ambulas [abt]
57. Alamblak [amp]
58. Yimas [yee]
59. Kiwai Southern [kjd]
60. Kewa [kew]
61. Kamasau [kms]
62. Meriam Mir [ulk]
63. Kobon [kpw]
64. Manam [mva]
65. Usan [wnu]
66. Tauya [tya]
67. Yagaria [qgr]
68. Hua [ygr]
69. Takia [tbc]
70. Waskia [wsk]
71. Menya [mcr]
72. Nabak [naf]
73. Kele [sbc]
74. Selepet [spl]
75. Koiari [kbk]
76. Yabem [jae]
77. Korafe [kpr]
78. Umanakaina [gdn]
79. Bali [bbn]
80. Mussau [emi]
81. Kove-Kaliai [kvc]
82. Gapapaiwa [pwg]
89. Kuot [kto]
90. Kol [kol]
91. Sulka [sua]
92. Madak [mmx]
93. Tolai [ksd]
94. Mali [gcc]
95. Duke of York [rai]
96. Siar [sjr]
97. Bandjalang [bdy]
98. Sudest [tgo]
99. Yélî Dnye [yle]
100. Halia [hla]
101. Rotokas [roo]
102. Banoni [bcm]
103. Motuna [siw]
104. Bilua [blb]
105. Sisiqa [qss]
106. Roviana [rug]
107. Lavukaleve [lvk]
108. Kokota [kkk]
109. Rennellese [mnv]
110. Longgu [lgu]
111. Cèmuhî [cam]
112. Xârâcùù [ane]
113. Aiwoo [nfl]
114. Iaai [iai]
115. Buma [tkw]
116. Mwotlap [mlv]
117. South Efate [erk]
118. Sye [erg]
119. Rotuman [rtm]
120. Fijian [fij]
121. Marquesan [mrq]

Among the conclusions to be drawn from the Reesink et al 2009 study are:

- Structural features of language can be used to help clarify historical relationships.
- In the study, large known groups of languages are recapitulated:

  - The Austronesian family with Oceanic as subgroup
  - The putative Trans New Guinea family, as proposed by Ross (2005), appeared as a solid block with the exception of the Alor-Pantar languages Klon and Abui and the Marind family (Marind and Inanwatan), separated from various non-TNG clusters
  - Australian languages are separated in Pama-Nyungan versus a non-PN cluster.
- However, some clusters represent hybridization rather than phylogeny, especially the cluster containing both Papuan and Austronesian languages of eastern Indonesia.

Some important questions remain: which features are responsible for the clustering? To what extent are structural features independent? Is it possible to distinguish phylogeny from lateral transfer? The issue of relative (in)dependence of structural features will be addressed in section 3 and in section 4 we will take a closer look at the hybrid cluster of eastern Indonesia identified above, applying the Structure algorithm to a new sample of Austronesian and Papuan languages of that area.

3. Relatively (in)dependent traits. After chance resemblance of features (due to the limited design space of language structure at the level of granularity that we have data for; see Dunn et al. 2005, Dunn et al. 2008 and Reesink et al. 2009), the main factors leading to resemblances between languages can be divided into two groups. Firstly, there are factors indicative of historical signal. These include shared inheritance from a common ancestral language, and diffusion through contact between speakers of different linguistic communities. Secondly there are factors which, while in some cases historically determined, do not allow us to infer individual language histories. These include logical entailment, typological dependency (implicational universals), and functionally motivated similarities due to system constraints (Croft 2008:230). For the purposes of making historical inferences about languages, this second set of factors acts as noise at best (obscuring a signal where present), and is misleading at worst (creating the appearance of a signal where one is absent). This is not to say that these factors are intrinsically bad for linguistic analysis: for making historical inferences about typological features this is exactly reversed. In an investigation of implicational universals shared history is the confound (see Dunn et al. 2011).

3.1. Establishing a set of structural features. For the original questionnaire used by Dunn et al. (2005), features were selected on the basis of what in the literature (Dunn et al. 2002; Foley 1998, 2000; Lynch et al. 2002) was known as typical or common characteristics of various Austronesian and Papuan lineages. Some improvements on that set was done for Dunn et al (2008:731), in part in response to commentary in Donohue and Musgrave (2007); see also Dunn et al. (2007). But at the start of the study reported in Reesink et al. (2009) we carried out a major overhaul of the questionnaire in consultation with colleagues (acknowledged in Reesink et al. 2009). Many questions were better defined, a number of questions were removed and others were added. In table 1 and table 2 we give some examples of original questions which could not easily be answered for many languages and which were replaced by questions whose terms were better defined

and more easily identified in a given description.

The questions whether there are adjectives and how they function attributively and predicatively caused some difficulties in the first version. This was solved by the new formulations, which specifically are meant to capture whether adjectival notions are nouny or verby in a particular language.[2]

	ADJECTIVES 2005/2008 [LANGUAGE]		ADJECTIVES 2006/2009 [PLOS BIOLOGY]
40	Is there lexical overlap between a significant proportion of adjectives and verbs (including zero-derivation)?	69	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) act like verbs in predicative position?
41	Does the same lexical set of adjectives function both attributively and predicatively?	70	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs?

Table 1. Questions relating to Adjectives in two versions

The original questionnaire used for Dunn et al. 2005 and 2008 contained a number of questions attempting to collect data on Tense-Aspect-Mood categories. Those questions were phrased in terms of "how many pure tenses are distinguished?" and "how many fused tense/mood categories are distinguished?" It was stipulated to " include affixes, clitics and satellite particles associated with verbs forming a constituent with the verb on some level, but exclude optional adverbials". Since the terms 'pure' versus 'fused' are not easily interpreted and because the answers were not binary as they are for all other traits, these questions weren't even used for those studies.

Thus, only the few questions in column 2 in table 2 were part of the analyses in the two studies, which meant that potentially important information regarding Tense marking could not be used. The revised questions in table 2 yield more clearly interpretable codes, and they restrict the traits to clearly morphological categories marked on the verb.

[2] As acknowledged in Reesink et al (2009:9), for comments and additions resulting in the latest version we thank Sjef Barbiers, Milly Crevels, Nick Evans, Rob Goedemans, Eva Lindström, Pieter Muysken, Gunter Senft, Leon Stassen, and Hein van der Voort (Workshop 15 May 2006, Radboud University and Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands). In particular the reformulation of the questions regarding adjectives is due to Leon Stassen.

	TAM 2005/2008 [LANGUAGE]		TAM 2006/2009 [PLOS BIOLOGY]
46	Do the same morphemes systematically encode both TAM and person?	79	Do verbs have prefixes/proclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)?
47	Do verbs have prefixes/proclitics?	80	Do verbs have suffixes/enclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)?
48	Do verbs have suffixes/enclitics?	81	Can infixation be used on verbs for derivational, aspectual, or voice-changing purposes?
		82	is there present tense regularly morphologically marked on the verb?
		83	is there past tense regularly morphologically marked on the verb?
		84	is there future tense regularly morphologically marked on the verb?
		85	are there multiple past or future tenses, distinguishing distance from Time of Reference, marked on the verb?
49	is a distinction between punctual/ continuous aspect available as a morphological choice?	86	is a distinction between punctual/ continuous aspect available as a morphological choice?
50	is a distinction between realis/ irrealis mood available as a morphological choice?	87	is a distinction between realis/irrealis mood available as a morphological choice?

TABLE 2. Features relating to Tense-Aspect-Mood affixation in two versions

For a full comparison of the differences between the two versions we refer to the Appendix. We continue with a discussion of the relative (in)dependence of traits in the most recent questionnaire.

3.2. LOGICAL ENTAILMENT. In spite of our attempt to minimize logical entailment between features in our database, there are some cases where we judge it innocuous to allow features with some degree of logical dependency between them to remain. For example, consider the possible values for two questions relating to the phonotactics of a language in (1):

(1) (a) Are there word-final consonants?

(b) Are there consonant clusters (not counting prenasalized consonants) in syllable coda?

The two questions are clearly not totally independent from each other, as particular values of certain features logically entail particular values of others:

if (a) = 1, then (b) = 1 or 0; if (a) = 0, then (b) = 0.
if (b) = 1, then (a) = 1; if (b) is 0, then (a) = 1 or 0.

However, the bias added by this dependency is small as this entailment is only partial, outweighed by the added statistical power we get from including data with the logically independent values. Given the large number of features in our analysis, it is not likely that this one case of partial dependency has seriously affected the results in our earlier analyses.

3.3. Chance resemblance due to limited design space. Most or all of the structural features of language have a far more restricted degree of freedom than lexical items. They are a fundamentally different kind of data with different statistical properties. For example, the two questions about the behavior of adjectival elements in predicative and attributive position (see table 1) were formulated to capture whether a language has verby (Y to both questions) or nouny (N to both) adjectives, or in between (Y to predicative; N to attributive verb-like behavior). A language which would have N to verb-like behavior in predicative position, but Y to verb-like behavior in attributive position was considered as unlikely. However, in our sample we do find this anomalous situation in the non-TNG language Imonda. Thus, the maximum number of four possibilities is available. This holds also for the two questions whether a language has prepositions or postpositions. There are languages with Y or N to both questions in addition to those that have only one or the other.

With regard to the order of Possessor and Possessum, the design space allows for three possibilities: the Possessor may (1) precede or (2) follow or (3) may do both. A negative value of both questions is of course not possible.

While such limited degrees of freedom may create homoplasies that do not reflect shared history, large-scale chance convergence is rendered unlikely through the use of a large number of features.

3.4. Typological dependency – implicational universals. Typological dependencies have been widely discussed since the sixties when Joseph Greenberg launched his language universals project. Most generalizations deal with word order properties in the clause and the nominal constituent. For example, it is well-known that OV order and postpositions are commonly found together, as are VO order and prepositions. Dunn et al (2011) has argued that there is a strong lineage-specific element to these apparent universals. Dryer (2005) presents data showing that the correlation is not perfect. Of a total of 1033 languages, 427 have OV and postpositions and 417 have VO and prepositions, while 10 languages combine OV with prepositions and 38 have VO together with postpositions. In addition, 141 languages do not fall into one of these four categories. For example, Dutch has prepositions but has both OV and VO order. On the other hand, Jabêm has SVO order with both prepositions and postpositions. Thus, while there is a strong typological tendency for the values of these features to be correlated, by removing some of these questions

important information is lost.

There are indeed rather high correlations between the questions on tense marking in table 2. However, combining past and future tense as reported by Dahl and Velupillai (2005; chapters 66 and 67 in WALS), there is no clear typological dependency cross-linguistically: of the 110 languages that mark future tense, there are 48 that mark a simple past tense, 26 with 2-3 degrees of remoteness, 1 with 4 or more degrees, and 35 with no past tense marking. In other words, if some of these questions were removed a considerable amount of information would be lost.

3.5. Functionally motivated – system constraints. Somewhat related to typological dependency is convergence due to system constraints. Some of our features may at first blush be mutually exclusive or inclusive. For instance, languages tend to have prepositions or postpositions, but relatively infrequently have both or neither. The raw counts for these features in our complete database (ignoring for the moment that these observations are phylogenetically dependent) are shown in table 3. The conditional probability of having prepositions given postpositions is 15%, and the conditional probability of having postpositions given prepositions is only 11%. A diachronic account for the development of adpositions predicts that the order of adposition and noun phrase will typically be fixed.

		Postpositions	
		Present	Absent
Prepositions	Present	11	87
	Absent	61	13

Table 3. Postpositions and prepositions.

Heine and Kuteva (2007) describe typical grammaticalization pathways such as *relational noun>adposition, adverb>adposition,* or *verb+complement>adposition+noun phrase,* which each have as their starting point a construction which most commonly already has fixed ordering. Even if two orders of adpositions and noun phrases are possible, the order will most likely be fixed with respect to the particular adposition selected. Given the constraint on adposition systems that there will usually be only one kind, there is a negative correlation between having prepositions and having postpositions.

Similar kinds of system constraints exist in other parts of the grammar. For example, there is a tendency for agreement affixes for transitive and intransitive subjects to be marked the same way. Thus, there is a positive correlation between having prefixes for marking transitive subjects (A) and intransitive subjects (S), and likewise there is a positive correlation between having suffixes for subjects of transitive (A) and subjects of intransitive clauses (S), as shown in table 4.

		S suffix	
		Present	Absent
A suffix	Present	11	4
	Absent	7	66
		S prefix	
		Present	Absent
A prefix	Present	35	3
	Absent	3	47
		O suffix	
		Present	Absent
O prefix	Present	1	0
	Absent	51	34

TABLE 4: A, S and O as prefixes and suffixes.

Some of these tendencies are nevertheless not strong. While there is a negative correlation between having object suffixes (O) and having object prefixes (see table 4), the amount of the variance this correlation explains of the (phylogenetically uncorrected) data is barely significant. This is despite a strong phylogenetic bias, in that no Austronesian languages have an object prefix, and most Trans New Guinea languages do. This would be expected to have the effect of exaggerating the apparent negative correlation between object prefixes and suffixes.

3.6 SHARED INHERITANCE. Correlations between features in a linguistic data set cannot be interpreted as causal with any validity without taking into account the confound introduced by possible genealogical relationships between the languages. This issue is known as Galton's problem: variables in languages related by common descent or diffusion are not statistically independent. Any apparent causal correlations between features of languages linked by shared history might be no more than 'duplicate copies of the same original' (Galton in Tylor 1889:270). This was alluded to above (section 3.4), with the example of object prefixes. Object prefixes are absent in Austronesian languages and highly frequent in Trans New Guinea languages. So, any other feature which is rare in Austronesian and common in TNG will correlate with the presence or absence of object prefixes. In a data set limited to Austronesian and TNG languages we could expect absurd correlations, such as positive correlations between object prefixes and altitude, negative correlations between object prefixes and navigational technology. These correlations are driven by accidents of history rather than any causal link.

In the full set of features we find substantial correlations, either positive or negative, which are clearly the product of shared history. For instance, there are positive correlations between Decimal counting systems, Prepositions, and the Inclusive/Exclusive distinction for non-singular first person. Likewise there is a negative correlation between these features

and verbal past tense.

Figure 2 illustrates the accidental nature of the negative correlation between decimal counting systems and verbal past tense marking. A decimal system of counting predominates in Austronesian, although there are quite a number of AN languages which exhibit a quinary system. And there is internal evidence in many of the Papuan languages with decimal systems that this occurred through contact with Austronesian speaking communities. The left panel of figure 2 shows a possible reconstruction of the history of decimal counting systems in a sample of Austronesian languages. The case for reconstructing decimal systems for proto-Austronesian seems strong. There are two sub-branches of languages lacking decimal counting systems and decimal counting systems occur on every level of the phylogeny. The right panel shows a somewhat different story. Verbal past tense marking occurs sporadically throughout the tree, but there are no cases where it makes sense to reconstruct past tense marking to an earlier node of the tree. The negative correlation between these features is apparently because of the relative stability of the two features, and their states in the ancestral languages. In other words, the clustering of such statistically dependent features is due to a genealogical signal.

The final possible cause of typological similarity between languages, diffusion through contact, will be discussed in section 4.

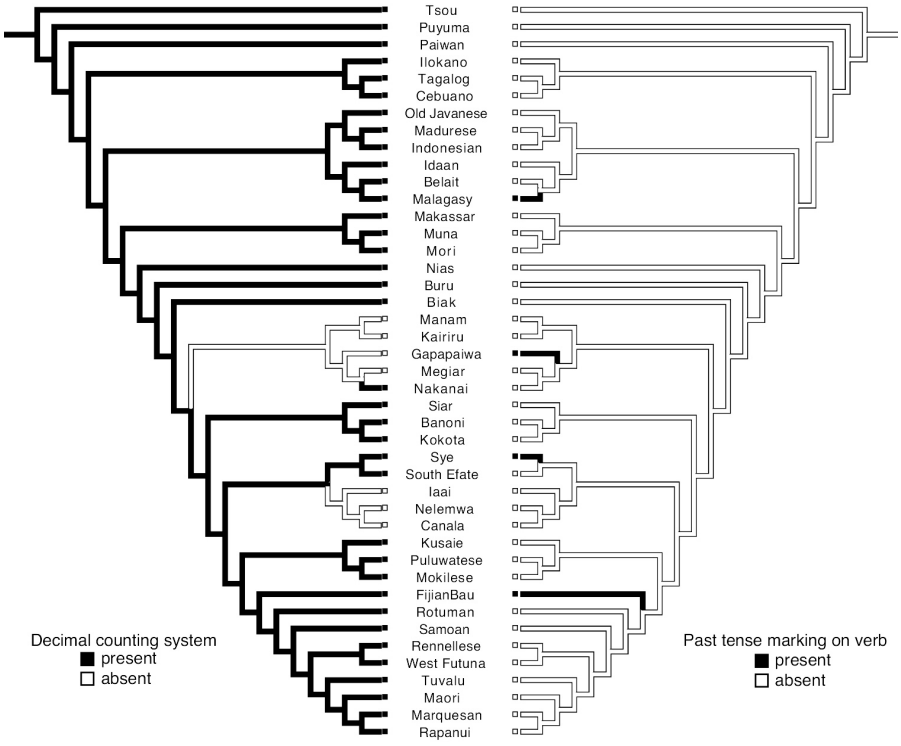

FIGURE 2. Decimal counting systems and past tense marking on the verb in Austronesian tree

4. Linguistic Populations in east Indonesia. The study examining linguistic traces of the Sahul Past (Reesink et al. 2009) employed the STRUCTURE algorithm (Pritchard et al. 2000), as shown in section 2. The method assumes a model in which there are a number (K) of unspecified or unknown populations, each of which is characterized by a set of allele frequencies at each locus. Individuals in any sample are assigned (probabilistically) to populations, or jointly to two or more populations if their genotypes indicate that they are admixed. The different values of the linguistic characters are the analogical equivalent of the genetic alleles, while a language is the equivalent of an individual in the biological studies. In other words, just as an individual's autosomal DNA is inherited from a number of different ancestors belonging to one or more biological populations, so a language may have inherited structural features from one or more different populations. The structure algorithm computes the most likely contribution of a given number (K) of ancestral populations to each of the individuals.

As stated at the end of section 2, Reesink et al. (2009) did find some striking correspondence between earlier defined linguistic families. However, as already mentioned, structural features cannot be used to claim or refute genealogical relationships between languages, see also Croft (2004). This is illustrated in the fact that the striking correspondence does not amount to full agreement among the groupings found by the different methods. A rather robust linguistic population identified by the Structure algorithm (Reesink et al. 2009) as the 'red' or 'West Papuan' cluster (see figure 1) contains all the Papuan languages of eastern Indonesia and the Bird's Head in the sample: Klon and Abui from the Alor-Pantar family, Tobelo and Tidore from North Halmahera, and Meyah and Hatam from the Bird's Head, as well as the two AN languages Taba and Biak. This cluster has also contributions to Papuan and Austronesian languages along the north coast of New Guinea and in the Bismarck archipelago. We concluded in that study: "This finding suggests an area of millennia of contact between AN and Papuan non-TNG speaking groups" (Reesink et al. 2009:8).

Given that earlier studies had shown a great degree of heterogeneity among the Papuan groups in east Indonesia (see for example Reesink 2005), it was rather surprising to see them clustered together with a few AN languages thrown in. Thus, new research questions are raised: 1) which features are responsible for a certain clustering; and 2) is it possible to differentiate phylogeny and diffusion?

In order to answer these questions a new study was conducted with two more AN languages from the same region, Tetun spoken in East Timor and Buru of the Moluccas, both classified as members of the Central Malayo-Polynesian subgroup. The validity of this subgroup, proposed by Blust as a linkage (1993), has been challenged by Donohue and Grimes (2008) and reaffirmed as most likely descending from a dialect chain by Blust (2009). We now report the results of this new study.

The Structure algorithm was applied this time to just a small sample of Austronesian languages of (eastern) Indonesia and Papuan languages of the same area. Since the algorithm simultaneously determines both the most likely number of ancestral groups and the most likely contribution of each of these populations to each of the observed individuals, we wanted to focus on the similarities and differences between just these languages, avoiding clustering that might ignore intragroup differences when compared to Papuan and Australian languages with different profiles, as was done in the major studies.

In figure 3 the clustering of these languages is shown for two to five ancestral populations (K2-5). For each specified number of clusters (K), the algorithm assigns a certain weight to each allele (in our case, the state of a particular feature). This clustering is independent for each K value, so that individuals may be assigned to different clusters (arbitrarily given a particular colour) on the basis of the amalgamated weights of the feature-states within each K.

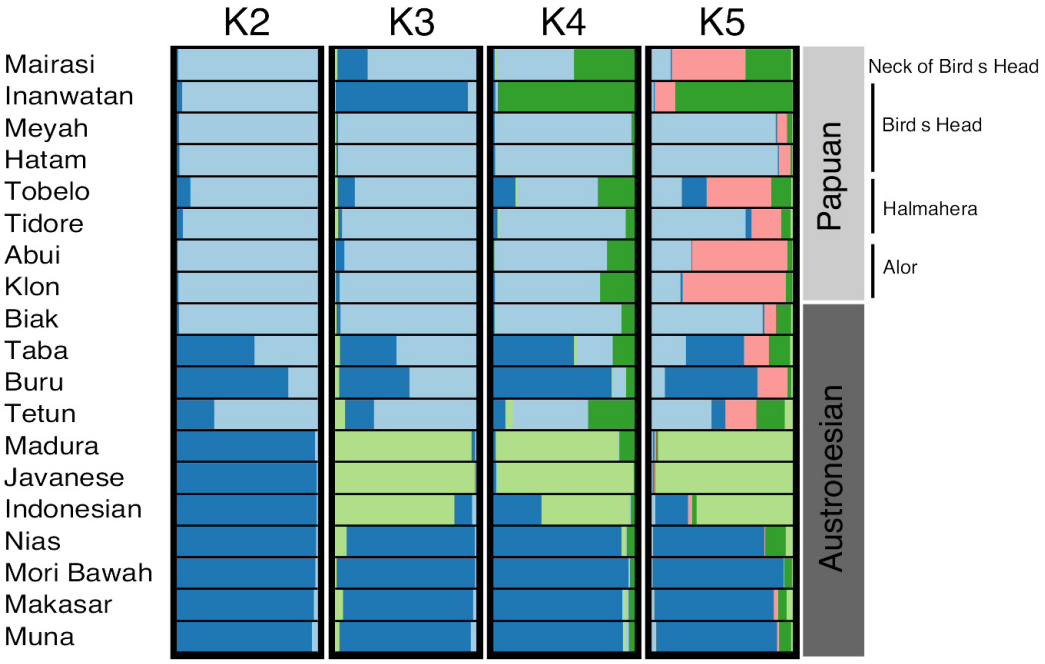

FIGURE 3. Clustering of AN and Papuan languages of eastern Indonesia

The K values 3, 4, and 5 hardly differ in their likelihood score. At K3 and K4 the light blue cluster contains AN and Papuan languages, while at K5 we find some differentiation. A new cluster (pink) is detected contributing mainly to Klon and Abui of the Alor-Pantar group, Tobelo of North Halmahera and Mairasi, spoken in the 'neck' connecting the Bird's Head to the rest of New Guinea. Thus at K5 we find a separation of a number of the Papuan languages, but still not all.

Clustering by the Structure algorithm is based on differential weighting to each of the 160 features per cluster. The same feature may have a higher or lower weighting for different K values. Space does not allow us to give a full list of different weights of each feature for each value of K, but in table 5 a sample pertaining to word order is given. These values show that presence of V final, Postpositions, and Object Prefix have a lower weight for the light blue cluster at K4 and thus, together with the values of other features, cannot differentiate Klon, Abui, Tobelo and Mairasi from the other Papuan and AN languages. At K5 these features have a stronger weighting, and thus a new cluster is identified.

Light blue	**K4**	**K5**	**Pink K5**
Verb final	0.38	0.24	0.61
Postposition	0.31	0.21	0.49
Object Prefix	0.38	0.22	0.56
Verb medial	0.64	0.79	0.45
Preposition	0.72	0.85	0.56

TABLE 5. Allele weights for features in contributing populations

Some values of the features in table 5 may look like a system constraint, or a typological correlation, but the overall correlation between Object prefix and Verb-final word order in the sample of 121 languages is rather weak (r = 0.40). The contribution of these features is therefore relatively independent.

In figure 3 it is clear that in all independent runs at all K values, the two unrelated Papuan languages of the Bird's Head, Hatam and Meyah, consistently cluster with the AN language Biak. It thus appears that in this case diffusion overrides phylogeny.

Is it possible to differentiate the two historical processes by extant structural features? We know from the comparison of their lexicons that Biak belongs to the South Halmahera-West New Guinea subgroup of the AN family and that Hatam and Meyah belong to two different Papuan families, albeit with perhaps a very remote common ancestor (Reesink 2002). Are there any traces in their structural features that still betray their genealogical affiliation? In other words, to what extent are these languages different in the set of structural features employed?

In order to find such traces we have to go into the nitty-gritty of the data. Table 6 lists all fifteen features (out of 160) on which the two unrelated Papuan languages Hatam and Meyah both agree with each other, presumably due to shared diffusion of Papuan traits, and are different in value from Austronesian Biak.

	Hatam	Meyah	Biak
Weight sensitive stress	-	-	+
Syllable position stress	-	-	+
Definite/specific articles	-	-	+
Indefinite article required	-	-	+
Difference comitative vs coordination	-	-	+
Gender in third person	-	-	+ (3pl.animate)
Numeral classifiers	+	+	-
Possession by suffix	-	-	+
Quinary counting system	+	+	-

	Hatam	Meyah	Biak
Attributive adjectives require same morphology as verbs	-	-	+
Copula for predicative N(P)	-	-	+
Aspectual auxiliaries	-	-	+
Causative by Serial Verb Construction	+	+	-
Nouns can be reduplicated	-	-	+
Other elements than N or V can be reduplicated	+	+	-

TABLE 6. Hatam and Meyah values agree and differ from Biak

These facts show very faint traces of structural features that may betray phylogenetic affiliation. For example, possession by suffix seems tightly linked to the AN family. In many AN languages to the north-west of this geographic region the Possessor normally follows the Possessum, and when that is expressed by a pronoun it can easily become encliticized or suffixed. This order is still present in Biak. It should be noted that the feature Possession by prefix (a separate question in our database) is not part of the list separating Hatam and Meyah from Biak, because for this trait all three languages have a positive value. This order is typical of the Papuan languages of the Bird's Head (and other regions of east Indonesia), and has diffused to a few AN languages in the Cenderawasih Bay area, in Biak and Ambai for plural possessors, in Waropen for both singular and plural (see Klamer et al. 2008:129). While all other Papuan languages of North Halmahera and the Bird's Head have a gender distinction for third person singular, the two east BH families that Hatam and Meyah belong to, do not. Yet Biak has adopted this Papuan trait in the form of a gender distinction between animate and inanimate for third person plural pronouns.

Of course, as mentioned above, single structural features can never be diagnostic for genealogical relatedness, and this is illustrated for these heterogeneous languages which have converged to such a degree that even their full structural profile obscures their descent. While the Comparative Method illuminates their genealogy, structural features illuminate their long term history of contact.

5. CONCLUSION. The results of large-scale comparison of structural features in a great number of languages from different lineages can be summarized as follows.

In population genetics the distribution and frequency of mutations in unrelated individuals are used to trace ancestral populations. In the studies reviewed in section 2 we practice *population linguistics*, that is, we attempt to find clusters between individual languages that are NOT immediate family. Where cognate-based methods cannot be applied, profiles of abstract structural features can discover plausible groupings in hitherto unrelated clusters of languages. These groupings may be the result of remote common ancestry, diffusion or both. In the case of a putative family like the Papuan TNG family, the result obtained by structural features may strengthen the tentative conclusions based on

pronominal forms. We do not claim that we now have conclusive evidence for TNG as a bona fide family, but simply that the proposed unity has some firmer footing.

Chance resemblances due to the limited degrees of freedom structural features have (Harrison 2003; section 3.3 above) can to some extent be overcome by considering a large number of features. Typological dependencies such as implicational universals and functionally motivated convergences are an empirical matter: how strong are they? They apparently differ in different lineages (Dunn et al. 2011).

The results reported in section 2 show that a large set of structural features does reveal a phylogenetic signal in that higher level linguistic groupings are identified. Due to their limited design space and relative ease of diffusion they cannot unequivocally identify lower level language families. As shown in section 4, the Structure algorithm cannot separate different lineages in eastern Indonesia, at least not with a strong likelihood. A matter for further research would be to investigate whether a different set of features could do better. It may be that a small set of diagnostic traits is masked by a much larger number of features that are shared by languages of different families by a Bayesian inference algorithm such as Structure, as illustrated for Hatam, Meyah and Biak in section 4.

While structural features can be diffused, complete substitution is quite rare. The basic morpho-syntactic profile, linked to the semantic-pragmatic way of representing the natural and social world of any particular speech community is quite robust through many descending generations. Therefore, the linguistic clusters found on the basis of full profiles provide information about their historical provenance. If it is possible to reconstruct/determine the ancestral state of a particular feature in a (putative) family, as for example shown by the presence of a decimal counting system and absence of past tense marking in the Austronesian family, then aberrant values in daughter languages can be accounted for by hybridization.

References

Blust, Robert. 1993. Central and Central-Eastern Malayo-Polynesian. *Oceanic Linguistics* 32(2). 241-293.

Blust, Robert. 2009. The position of the languages of eastern Indonesia: A reply to Donohue and Grimes. *Oceanic Linguistics* 48(1). 6-77.

Croft, William. 2004. Typological traits and genetic linguistics. http://www.unm.edu/~wcroft/Papers/Typ-Gen.pdf. (5 November, 2010.)

Croft, William. 2008. Evolutionary linguistics. *Annual Review of Anthropology* 37. 219–234.

Dahl, Östen & Velupillai, Viveka. 2011. The Past Tense. In Dryer & Haspelmath, http://wals.info/chapter/66. (8 September, 2012.)

Dahl, Östen & Velupillai, Viveka. 2011. The Future Tense. In Dryer & Haspelmath, http://wals.info/chapter/67. (8 September, 2012.)

Donohue, Mark, & Simon Musgrave. 2007. Typology and the linguistic macrohistory of Island Melanesia. *Oceanic Linguistics* 46(2). 325–364.

Donohue, Mark, & Charles E. Grimes. 2008. Yet more on the position of the languages of Eastern Indonesia and East Timor. *Oceanic Linguistics* 47(1). 114-158.

Dryer, Matthew S. 2011. Relationship between the order of object and verb and the order of adposition and noun phrase. In Dryer & Haspelmath, http://wals.info/chapter/95. (10 September, 2012.)

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The world atlas of language structures online*. http://wals.info/.

Dunn, Michael, Ger Reesink & Angela Terrill. 2002. The East Papuan languages: A preliminary typological appraisal. *Oceanic Linguistics* 41(1). 28–62.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309. 2072–2075.

Dunn, Michael, Robert A. Foley, Stephen C. Levinson, Ger Reesink & Angela Terrill. 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46(2). 388-403.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, & Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84(4). 710-759.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82.

Foley, William A. 1998. Toward understanding Papuan languages. In Miedema et al., 503–18.

Foley, William A. 2000. The languages of New Guinea. *Annual Review of Anthropology* 29. 357-404.

Greenberg, Joseph H 1971. The Indo-Pacific hypothesis. In Thomas A. Sebeok (ed.), *Linguistics in Oceania* (Current Trends in Linguistics 8), 807–71. The Hague: Mouton.

Harrison, S. P. 2003. On the limits of the comparative method. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 213–43. London: Blackwell.

Heine, Bernd & Tania Kuteva. 2007. *The genesis of grammar*. Oxford: Oxford University Press.

Klamer, Marian, Ger Reesink & Miriam van Staden. 2008. Eastern Indonesia as a linguistic area. In Pieter Muysken (ed.) *From linguistic areas to areal linguistics*, 95-149. Amsterdam: Benjamins.

Lynch, John, Malcolm Ross & Terry Crowley. 2002. *The Oceanic Languages*. London: Curzon.

Miedema, Jelle, Cecilia Odé & Rien A.C. Dam (eds.). 1998. *Perspectives on the Bird's Head of Irian Jaya.* Amsterdam: Rodopi.

Pawley, Andrew K. 1998. The Trans New Guinea phylum hypothesis: A reassessment. In Miedema et al., 655-690.

Pawley, Andrew. 2005. The chequered career of the trans New Guinea hypothesis: Recent research and its implications. In Pawley et al., 67-107.

Pawley, Andrew, Robert Attenborough, Jack Golson & Robin Hide (eds.). 2005. *Papuan Pasts, Studies in the cultural, linguistic and biological history of the Papuan-speaking peoples*. Canberra: Pacific Linguistics.

Pritchard, Jonathan, Matthew Stephens, & Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155. 945–959.

Reesink, Ger. 2002. *Languages of the eastern Bird's Head*. Canberra: Pacific Linguistics.
Reesink, Ger. 2005. West Papuan languages: Roots and development. In Pawley et al., 185-218.
Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population methods. *PloS Biology* 7(11). e1000241.
Ross, Malcolm. 2001. Is there an East Papuan phylum? Evidence from pronouns. In Andrew Pawley, Malcolm Ross and Darrell Tryon (eds.), *The boy from Bundaberg: Studies in Melanesian linguistics in honour of Tom Dutton,* 301-321. Canberra: Pacific Linguistics.
Ross, Malcolm. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. In Pawley et al., 15-65.
Wurm, Stephen A. 1975. The East Papuan phylum in general. In Stephen A. Wurm (ed.), *Papuan languages and the New Guinea linguistic scene*, 783-803. Canberra: Pacific Linguistics.
Wurm, Stephen A. 1982. *Papuan languages of Oceania* (Ars Linguistica 7). Tübingen: Gunter Narr.

Ger Reesink
ger.reesink@hccnet.nl

Michael Dunn
michael.dunn@mpi.nl

Appendix

In this table the lists of characters used for Dunn et al. (2008) in *Language* and for Reesink et al. (2009) in *PloS Biology* are compared.

	characters 'Language 2008'		**characters 'PloS 2009'**
		1	Are there as many points of articulation for nasals as there are for stops? *Only consider points of articulation where a nasal is phonetically possible* (1: present, 0: absent)
		2	Is there contrast between heterorganic and homorganic sequence of nasal and velar stop? *For example, does the language permit a phonetic contrast between -nk- and -ŋk- clusters* (1: present, 0: absent)
1	Are there fricative phonemes?	3	Are there fricative phonemes? (1: present, 0: absent)
2	Are there phonemic prenasalised stops?	4	Are there phonemic prenasalised stops? (1: present, 0: absent)
3	Is there a phonemic distinction between l/r?	5	Is there a phonemic distinction between l/r? (1: present, 0: absent)
4	Is there a phonemic velar fricative or glide?	6	Is there a phonemic velar fricative or glide? (1: present, 0: absent)
5	Is there a voicing contrast between oral (i.e. non-prenasal) stops?	7	Is there a voicing contrast between oral (i.e. non-prenasal) stops? (1: present, 0: absent)
		8	Is there a laminal/apical contrast? (1: present, 0: absent)
		9	Are there retroflexed consonants? (1: present, 0: absent)
6	Is there phonemic consonant length?	10	Is there phonemic consonant length? (1: present, 0: absent)
7	Is there phonemic vowel length?	11	Is there phonemic vowel length? (1: present, 0: absent)
8	Are there contrastive phonation types for vowels? (e.g. nasal, creaky, etc)		

	characters 'Language 2008'		**characters 'PloS 2009'**
		12	Are there two or more contrastive central vowels *Do not include length contrasts* (1: present, 0: absent)
9	Is there lexically determined suprasegmental prominence? *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena (don't include phonemic vowel length)*	13	Is there lexically determined suprasegmental prominence? *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena (don't include phonemic vowel length)* (1: present, 0: absent)
		14	Is there weight-sensitive suprasegmental prominence *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena* (1: present, 0: absent)
		15	Is there syllable position sensitive suprasegmental prominence? *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena* (1: present, 0: absent)
		16	Is there a tonal system? *I.e. two or more contrastive tones* (1: present, 0: absent)
10	Are there word-final consonants?	17	Are there word-final consonants? (1: present, 0: absent)
11	Are there consonant clusters?	18	Are there consonant clusters (not counting prenasalized consonants) in syllable onset? (1: present, 0: absent)
		19	Are there consonant clusters (not counting prenasalized consonants) in syllable coda? (1: present, 0: absent)
12	Are there definite or specific articles?	20	Are there definite or specific articles? (1: present, 0: absent)
13	Are there indefinite or non-specific articles?	21	Is an indefinite NP obligatorily accompanied by an indefinite (or non-specific) article? *Disregard if only on personal names* (1: present, 0: absent)
14	Is the order of NP elements Art N?	22	Are there prenominal articles? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
15	Are NPs N-initial (except for articles)?	23	Are there postnominal articles? (1: present, 0: absent)
		24	What is the relative position of numeral and noun in the NP? (*multistate* 1; Num-N; 2: N-Num; 3: both.)
		25	What is the relative position of demonstrative and noun in the NP? (*multistate* 1: Dem-N; 2: N-Dem; 3: both.)
		26	Are there 'discontinuous noun phrases'? *Can an argument be expressed by multiple N/NP throughout the clause > i.e. the Australian type.* (1: present, 0: absent)
		27	Is there a difference between the marking of NP coordination ('John and Mary went to market') and the marking of comitative phrases ('John went to market with Mary')? (1: present, 0: absent)
16	Is there an inclusive/exclusive distinction?	28	Is there an inclusive/exclusive distinction? (1: present, 0: absent)
		29	Is there a minimal-augmented system? *i.e. four basic pronominal forms for 1sg, 2sg, 3sg and 1+2, which each can be affixed for plural (or dual etc.)* (1: present, 0: absent)
		30	Is there a gender distinction in 3rd person pronouns (or demonstratives, if no 3rd person pronouns)? *either two- or threefold* (1: present, 0: absent)
		31	Is there a dual (or unit augmented) in addition to a plural (or augmented) number category in pronouns? (1: present, 0: absent)
17	Are 1st and 2nd persons conflated in any context?	32	Are 1st and 2nd persons conflated in any context? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
18	Are 2nd and 3rd persons conflated in non-singular numbers? (*Morphologically in any paradigm. Disregard pragmatics/politeness*)	33	Are 2nd and 3rd persons conflated in non-singular numbers? *morphologically in any paradigm. Disregard pragmatics/politeness* (1: present, 0: absent)
		34	Are person categories neutralized under some conditions? *e.g. in non-singular, under NEG, in certain TAM* (1: present, 0: absent)
19	Are more than 2 degrees of distance morphologically marked in demonstratives?	35	Is there an opposition between three or more distance terms in the demonstrative system? (1: present, 0: absent)
20	Are any of the spatial demonstratives not speaker-based? *Speaker-based spatial demonstratives are demonstratives that take as their decitic centre the speaker. By contrast, some demonstratives take not the speaker but the addressee as the deictic centre, for example a demonstrative might mean 'close to the speaker'; and some take both speaker and addressee as the deictic centre e.g. 'far from speaker and addressee'.*		
21	Is elevation morphologically marked in demonstratives?	36	Is elevation morphologically marked in demonstratives? (1: present, 0: absent)
		37	Is the opposition visible-non-visible marked on demonstratives? (1: present, 0: absent)
22	Are demonstratives classified?	38	Are demonstratives classified? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
23	Are there declensions (partly) determined by number of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts*	39	Are there declensions (partly) determined by number of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts* (1: present, 0: absent)
24	Are there declensions (partly) determined by gender of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts*	40	Are there declensions (partly) determined by gender of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts* (1: present, 0: absent)
25	Are there nouns which are suppletive for number? (*Only yes if present for more than **2** (basic) kin terms*)	41	Are there nouns which are suppletive for number? *Only answer yes if present for more than 2 (basic) kin terms*
		42	Can singular number be marked on the noun itself? *Number marking on noun does not count phrase level clitic or reduplication; absence of plural marking does not count as singular marking; exclude derivational forms (e.g. deverbal, deadjectival)* (1: present, 0: absent)
26	Can dual number be marked on the noun itself? *Number-marking on N does not count phrase-level clitic or reduplication*	43	Can dual number be marked on the noun itself? *number-marking on N does not count phrase-level clitic or reduplication* (1: present, 0: absent)
		44	Can plural number be marked on the noun itself? *number-marking on N does not count phrase-level clitic or reduplication* (1: present, 0: absent)
27	Is number marking prohibited on certain (types of) nouns? *(do not include proper nouns, e.g. place names or personal names)*	45	Is number marking prohibited on certain (types of) nouns? *(do not include proper nouns, e.g. place names or personal names)* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		46	Are there associative plurals? *e.g. Mary-PL = Mary and her family* (1: present, 0: absent)
		47	Is there a productive morphologically marked Action/state nominalization (arrive-arrival)? *if a language is precategorial, include the morphological mechanisms to produce such 'nominalizations'* (1: present, 0: absent)
		48	Is there a productive morphologically marked Agentive nominalization (sing-er)? (1: present, 0: absent)
		49	Is there a productive morphologically marked Object nominalization (sing; song)? (1: present, 0: absent)
28	Are there noun classes/genders? *By noun classes/genders is meant a system of dividing all or almost all of the nouns of a language into morphological classes which determine agreement phenomena beyond the noun itself.*	50	Are there noun classes/genders? *By noun classes/genders is meant a system of dividing all or almost all of the nouns of a language into morphological classes which determine agreement phenomena beyond the noun itself.* (1: present, 0: absent)
		51	Is sex a relevant category in the noun class/gender system? (1: present, 0: absent)
		52	Is shape a relevant category in the noun class/gender system? (1: present, 0: absent)
		53	Is animacy (without reference to sex) a relevant category in the noun class/gender system? (1: present, 0: absent)
		54	Is plant status a relevant category in the noun class/gender system? (1: present, 0: absent)
		55	Does the language only have a gender distinction in 3rd person pronouns? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		56	Is there concord within the NP, i.e. agreement of elements within the NP with the noun class of a noun? *related to class/gender* (1: present, 0: absent)
29	Are there numeral classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorization devices, the choice of which is determined by lexical selection*	57	Are there numeral classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorisation devices, the choice of which is determined by lexical selection* (1: present, 0: absent)
30	Are there possessive classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorisation devices, the choice of which is determined by lexical selection*	58	Are there possessive classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorisation devices, the choice of which is determined by lexical selection* (1: present, 0: absent)
31	Are there possessive classes? *i.e. different nouns treated differently in possession according to semantically-based groupings. Include alienable/inalienable*.		
32	Is alienable/inalienable a relevant distinction?	59	Is alienable/inalienable a relevant distinction? (1: present, 0: absent)
33	Are there different possessive constructions?	60	Are there different possessive constructions? (1: present, 0: absent)
		61	Can possession be marked by a prefix? *even if only on a restricted numer of kin terms. Emphasis is on *can** (1: present, 0: absent)
		62	Can possession be marked by a suffix? *even if only on a restricted numer of kin terms. Emphasis is on *can** (1: present, 0: absent)
34	Can possession be marked on the nominal possessor?	63	Can possession be marked on the nominal possessor? (1: present, 0: absent)
35	Can possession be marked on the nominal possessee?	64	Can possession be marked on the nominal possessee? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
36	If the order of elements in a possessive construction is fixed, is it possessor-possessed?	65	What is the relative position of possessor and possessed in the attributive possessive construction? (*multistate* 1:Possessor-Possessed; 2:Possessed-Possessor; 3: both)
		66	Are there different orders of elements in a possessive phrase for different classes of possession? *emphasis on *for different types of possession** (1: present, 0: absent)
37	Is there a decimal counting system? (*i.e. elements of decimal; even lexical **10**, **10**+**5** qualify*.)	67	What is the counting system? (*multistate* 1:Decimal; 2:Quinary; 3: Body-part tallying; 4: minimal) [Other systems, like senary, are not scored]
38	Is there evidence for any element of a quinary counting system? (*e.g. expressions for 5+1, 10+5+1*.)		
39	Are there words for particular amounts of a thing? *(e.g. ten possums)*		
40	Is there lexical overlap between a significant proportion of adjectives and verbs (including zero-derivation)?	68	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) act like verbs in predicative position? (1: present, 0: absent)
41	Does the same lexical set of adjectives function both attributively and predicatively?	69	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs? (1: present, 0: absent)
42	Is there case marking for core nominal NPs (i.e., S, A or O function)? *For case marking, include any affixal marking which appears in the NP and shows the function of the NP in the clause; adpositions are not counted.*	70	Is there case marking for core nominal NPs (i.e., S, A or O function)? *for case marking, include any affixal marking which appears in the NP and shows the function of the NP in the clause; do not count adpositions* (1: present, 0: absent)
		71	Is there case marking for core pronouns? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
43	Is there case marking for oblique nominal NPs ? *e.g. locationals, instrumentals, etc.; adpositions are not counted.*	72	Is there case marking for oblique nominal NPs ? *e.g. locationals, instrumentals, etc. do not count adpositions* (1: present, 0: absent)
		73	Is there case marking for oblique pronouns? (1: present, 0: absent)
44	Are there prepositions?	74	Are there prepositions? (1: present, 0: absent)
45	Are there postpositions?	75	Are there postpositions? (1: present, 0: absent)
		76	Are there adpositions to mark core NPs? (1: present, 0: absent)
		77	Are there adpositions to mark oblique NPs? (1: present, 0: absent)
		78	Is there a distinction between locational and directional adpositions? (1: present, 0: absent)
46	Do the same morphemes systematically encode both TAM and person?	79	Do verbs have prefixes/proclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)? *A, S, and O affixes are dealt with in 3.3* (1: present, 0: absent)
47	Do verbs have prefixes/proclitics?		
48	Do verbs have suffixes/enclitics?		
		80	Do verbs have suffixes/enclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)? (1: present, 0: absent)
		81	Can infixation be used on verbs for derivational, aspectual, or voice-changing purposes? (1: present, 0: absent)
		82	Is there present tense regularly morphologically marked on the verb? (1: present, 0: absent)
		83	Is there past tense regularly morphologically marked on the verb? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		84	Is there future tense regularly morphologically marked on the verb? (1: present, 0: absent)
		85	Are there multiple past or future tenses, distinguishing distance from Time of Reference, marked on the verb? (1: present, 0: absent)
49	Is a distinction between punctual/ continuous aspect available as a morphological choice?	86	Is a distinction between punctual/ continuous aspect available as a morphological choice? (1: present, 0: absent)
50	Is a distinction between realis/irrealis mood available as a morphological choice?	87	Is a distinction between realis/irrealis mood available as a morphological choice? (1: present, 0: absent)
		88	Is there an apprehensive modal category marked on the verb *also known as 'evitative', 'lest', etc* (1: present, 0: absent)
51	Is the S participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology*	89	Is the S participant (at least sometimes) marked by a suffix/enclitic? *pertains to verb morphology* (1: present, 0: absent)
52	Is the S participant (at least sometimes) marked by a prefix/ proclitic? *pertains to verb morphology*	90	Is the S participant (at least sometimes) marked by a prefix/proclitic? *pertains to verb morphology* (1: present, 0: absent)
53	Is the A participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology*	91	Is the A participant (at least sometimes) marked by a suffix/enclitic? *pertains to verb morphology* (1: present, 0: absent)
54	Is the A participant (at least sometimes) marked by a prefix/ proclitic? *pertains to verb morphology*	92	Is the A participant (at least sometimes) marked by a prefix/proclitic? *pertains to verb morphology* (1: present, 0: absent)
55	Is the O participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology*	93	Is the O participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
56	Is the O participant (at least sometimes) marked by a prefix/ proclitic? *pertains to verb morphology*	94	Is the O participant (at least sometimes) marked by a prefix/proclitic? *pertains to verb morphology* (1: present, 0: absent)
57	Are variations in marking strategies of core participants based on TAM distinctions?	95	Are variations in marking strategies of core participants based on TAM distinctions? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
58	Are variations in marking strategies based on verb classes?	96	Are variations in marking strategies based on verb classes? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
59	Are variations in marking strategies based on clause type, e.g. main vs subordinate?	97	Are variations in marking strategies based on clause type, e.g. main vs subordinate? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
60	Are variations in marking strategies based on person distinctions?	98	Are variations in marking strategies based on person distinctions? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
61	Do verb stems alter according to the **number** of a core participant?		
62	Do verb stems alter according to the **person** of a core participant?	99	Do verb stems alter according to the person of a core participant? (1: present, 0: absent)
63	Is number ever marked separately from person on the verb?	100	Is number ever marked separately from person on the verb? (1: present, 0: absent)
64	Are person, number and any TAM category (i.e. **3** or more categories in all) marked by portmanteau morphemes on verbs?	101	Are person, number and any TAM category (i.e. 3 or more categories in all) marked by portmanteau morphemes on verbs? (1: present, 0: absent)
65	Are categories such as person, number, gender related to a single participant discontinuously marked on a verb?	102	Are categories such as person, number, gender related to a single participant discontinuously marked on a verb? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
66	Is a non-core participant marked on the verb? *Include affixes, clitics and satellite particles associated with verbs forming a constituent with the verb on some level, but exclude optional adverbials*.	103	Are benefactive nominals marked on the verb? (1: present, 0: absent)
		104	Can instruments be marked on the verb? (1: present, 0: absent)
67	Can recipients be treated as a transitive object, i.e. as Direct object?	105	Can recipients be treated as a transitive object, i.e. as Direct Object? (1: present, 0: absent)
68	Are there syntactically ditransitive verbs?	106	Are there syntactically ditransitive verbs? (1: present, 0: absent)
69	Is negation marked morphologically on the verbs? *i.e. affixation, stem alternation, neutralization of some inflection*	107	Is negation marked morphologically on the verbs? *i.e. affixation, stem alternation, neutralization of some inflection* (1: present, 0: absent)
70	Is direction marked on verbs *Includes affixes, clitics and satellite particles associated with verbs forming a constituent with the verb on some level, but excludes optional adverbials.*	108	Can locative or direction be morphologically marked on the verb? *Locative as Direct Object ('she sleeps mat') does not qualify* (1: present, 0: absent)
71	Are there suppletive verbs for number of participants	109	Are there suppletive verbs for number of participants? *(list them all if feasible, otherwise give an estimate of the number and/or proportion of nouns)* (1: present, 0: absent)
		110	Are there suppletive verbs for tense or aspect? (1: present, 0: absent)
72	Are there conjugation classes?	111	Are there conjugation classes? (1: present, 0: absent)
73	Are there (several) verbs which can be used either transitively or intransitively with no morphological marking? *say no if it's only one or two stems; Intended here is the 'break' and 'open' type; not John eats/ eats the bread*	112	Are there (several) verbs which can be used either transitively or intransitively with no morphological marking? *say no if it's only one or two stems; Intended here is the 'break' and 'open' type; not John eats/ eats the bread* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
74	Is there transitivising morphology (include clitics)?	113	Is there transitivising morphology (include clitics)? (1: present, 0: absent)
75	Is there morphology (include clitics) to mark a reflexive action? *free word/ particle does not count; neither a default P/N co-reference*	114	Is there morphology (include clitics) to mark a reflexive action? *free word/ particle does not count; neither a default P/N co-reference* (1: present, 0: absent)
76	Is there morphology (include clitics) to mark a reciprocal action? *free word/particle does not count; neither a default P/N co-reference*	115	Is there morphology (include clitics) to mark a reciprocal action? *free word/particle does not count; neither a default P/N co-reference* (1: present, 0: absent)
77	Do verbs classify the shape, size, consistency or position of absolutive arguments by means of incorporated nouns, verbal affixes or suppletive verb stems? *not included here are positional verbs that classify a referent in such terms*	116	Do verbs classify the shape, size, consistency or position of absolutive arguments by means of incorporated nouns, verbal affixes or suppletive verb stems? *not included here are positional verbs that classify a referent in such terms - covered by 127* (1: present, 0: absent)
78	Is there a copula for predicate nouns? *e.g. John is a teacher*	117	Is there a copula for predicate nouns? *e.g. John is a teacher* (1: present, 0: absent)
79	Are there serial verb constructions? *(i.e. two or more verbs in juxtaposition, functioning as a single predicate, with no morphology to mark their relationship with each other. Each of the verbs is a separate phonological word but the construction as a whole is expressed in one intonational unit. Morphology is shared to a greater or lesser extent.)*	118	Are there serial verb constructions? *(i.e. two or more verbs in juxtaposition, functioning as a single predicate, with no morphology to mark their relationship with each other. Each of the verbs is a separate phonological word but the construction as a whole is expressed in one intonational unit. Morphology is shared to a greater or lesser extent.)* (1: present, 0: absent)
80	Is there one or more auxiliary?		
		119	Are there modal auxiliaries? (1: present, 0: absent)
		120	Are there aspectual auxiliaries? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		121	Are there tense auxiliaries? (1: present, 0: absent)
81	Is verb compounding a regular process? *(i.e. two or more verb stems acting as one phonological and grammatical word)*	122	Is verb compounding a regular process? *(i.e. two or more verb stems acting as one phonological and grammatical word)* (1: present, 0: absent)
82	Are there verb-adjunct (aka light-verb) constructions? *(i.e. constructions involving a non-predicating element expressing the lexical meaning of the construction, in conjunction with a semantically fairly empty verb, which enables the element to function as a predicate by providing the necessary morphology, e.g. eye do for 'see'; or sneeze hit for 'sneeze')*	123	Are there verb-adjunct (aka light-verb) constructions? *(i.e. constructions involving a non-predicating element expressing the lexical meaning of the construction, in conjunction with a semantically fairly empty verb, which enables the element to function as a predicate by providing the necessary morphology, e.g. eye do for 'see'; or sneeze hit for 'sneeze')* (1: present, 0: absent)
83	Is there incorporation of any element into verbs?		
		124	Is there incorporation of nouns into verbs a productive intransitivizing process? (1: present, 0: absent)
		125	Is there productive incorporation of other elements (adjectives, locatives, etc.) into verbs? (1: present, 0: absent)
84	Is there one or more existential verb? *exclude e.g. positional verbs*	126	Is there one or more existential verb? *exclude e.g. positional verbs (3.8.02)* (1: present, 0: absent)
		127	Are there positional (classificatory) verbs? *(i.e. in answer to a question 'Where is the X', does the verb used in the answer depend on the type of referent (e.g. do you have to say 'The X sits/stands/lies/etc on the table'). List them all.* (1: present, 0: absent)
85	Is the verb 'give' morphologically peculiar (different from most other verbs)? *e.g. stem suppletion, different affixation*	128	Is the verb 'give' morphologically peculiar (different from most other verbs)? *e.g. stem suppletion, different affixation* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
86	Is there a notably small number, i.e. about 100 or less, of verbs in the language?	129	Is there a notably small number, i.e. about 100 or less, of verbs in the language? (1: present, 0: absent)
87	Is a pragmatically unmarked constituent order SV for intransitive clauses?	130	What is the pragmatically unmarked order of S and V in instransitive clauses? (*multistate* 1: SV; 2: VS; 3: both)
88	Is a pragmatically unmarked constituent order VS for intransitive clauses?		
89	Is a pragmatically unmarked constituent order verb-initial for transitive clauses?	131	Is a pragmatically unmarked constituent order verb-initial for transitive clauses? (1: present, 0: absent)
90	Is a pragmatically unmarked constituent order verb-medial for transitive clauses?	132	Is a pragmatically unmarked constituent order verb-medial for transitive clauses? (1: present, 0: absent)
91	Is a pragmatically unmarked constituent order verb-final for transitive clauses?	133	Is a pragmatically unmarked constituent order verb-final for transitive clauses? (1: present, 0: absent)
		134	Is the order of constituents the same in main and subordinate clauses? (1: present, 0: absent)
		135	Do clausal objects occur in the same position as nominal objects? (1: present, 0: absent)
92	Is constituent order fixed? *Do not consider 'left or right-dislocation', accompanied by intonational signals*	136	Is constituent order fixed? *Do not consider 'left or right-dislocation', accompanied by intonational signals* (1: present, 0: absent)
93	Can negation be marked clause-finally? *This includes suffixes on verb-final clauses; prefixes on clause-final verbs do not count; Don't include elliptical 'Pete didn't'*	137	Can negation be marked clause-finally? *This includes suffixes on verb-final clauses; prefixes on clause-final verbs do not count; Don't include elliptical 'Pete didn't'* (1: present, 0: absent)
94	Can negation be marked clause-initially? *Don't include elliptical 'Not Mary'*	138	Can negation be marked clause-initially? *Don't include elliptical 'Not Mary'* (1: present, 0: absent)

	characters 'Language 2008'	**characters 'PloS 2009'**
95	Is there a difference between imperative and declarative negation?	139 Is there a difference between imperative and declarative negation? (1: present, 0: absent)
96	Are verbal and non-verbal predicates marked by the same negator?	140 Are verbal and non-verbal predicates marked by the same negator? (1: present, 0: absent)
97	Are S and O conflated morphologically in at least some basic constructions, i.e. simple main clauses?	141 Are S and O conflated morphologically in at least some basic constructions, i.e. simple main clauses? (1: present, 0: absent)
98	Are S and A conflated morphologically in at least some basic constructions, i.e. simple main clauses?	142 Are S and A conflated morphologically in at least some basic constructions, i.e. simple main clauses? (1: present, 0: absent)
99	Are S and O conflated morphologically across clause boundaries, i.e. acting as syntactic pivot?	143 Are S and O conflated morphologically across clause boundaries, i.e. acting as syntactic pivot? (1: present, 0: absent)
100	Are S and A conflated morphologically across clause boundaries, i.e. acting as syntactic pivot?	144 Are S and A conflated morphologically across clause boundaries, i.e. acting as syntactic pivot? (1: present, 0: absent)
101	Do S and O operate in the same way, and differently from A, for the purpose of any syntactic construction?	145 Do S and O operate in the same way, and differently from A, for the purpose of any syntactic construction? (1: present, 0: absent)
102	Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states?	146 Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states? (1: present, 0: absent)
		147 Is there a morphologically marked passive construction? *morphological marking includes some verbal affixation or some periphrastic element in the VP or clause* (1: present, 0: absent)
		148 Is there a morphologically marked antipassive? *morphological marking includes some verbal affixation or some periphrastic element in the VP or clause* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		149	Is there a morphologically marked inverse? *i.e. different marking by verbal affixation or pronominal clitics referring to A and O, depending on person, animacy or definiteness* (1: present, 0: absent)
103	Is there clause chaining? *i.e. chains of morphologically stripped-down medial clauses which are dependent on a single clause (usually, but not necessarily, final) for their TAM or participant marking specification*	150	Is there clause chaining? *i.e. chains of morphologically stripped-down medial clauses which are dependent on a single clause (usually, but not necessarily, final) for their TAM or participant marking specification* (1: present, 0: absent)
		151	Is there a morphologically-marked switch reference system? (1: present, 0: absent)
104	Is there a morphologically marked distinction between simultaneous and sequential clauses?	152	Is there a morphologically marked distinction between simultaneous and sequential clauses? (1: present, 0: absent)
105	Is the verb 'say' or a quotative construction used in desiderative constructions? *(e.g. 'I said for him to go' for 'I wanted him to go')*	153	Is the verb 'say' or a quotative construction used in desiderative constructions? *(e.g. 'I said for him to go' for 'I wanted him to go')* (1: present, 0: absent)
106	Are there purposive non-finite subordinate clauses?		
107	Are there temporal non-finite subordinate clauses?		
108	Are there complement clauses?		
109	Are causatives formed by serial verb constructions?	154	Are causatives formed by serial verb constructions? (1: present, 0: absent)
110	Are causatives formed by bound affixes/clitics?	155	Are causatives formed by bound affixes/clitics? (1: present, 0: absent)
111	Are causatives formed by constructions involving 'say'?	156	Are causatives formed by constructions involving 'say'? (1: present, 0: absent)
112	Is topic or focus marked morphologically? *i.e. by affixes or clitics*.		

	characters 'Language 2008'		**characters 'PloS 2009'**
113	Is there tail-head linkage? *(i.e. a discourse strategy in which the final verb of one sentence is repeated as the first verb of the next sentence)*	157	Is there tail-head linkage? *(i.e. a discourse strategy in which the final verb of one sentence is repeated as the first verb of the next sentence)* (1: present, 0: absent)
114	Are verbs reduplicated?	158	Are verbs reduplicated? (1: present, 0: absent)
115	Are nouns reduplicated?	159	Are nouns reduplicated? (1: present, 0: absent)
		160	Are elements apart from verbs or nouns reduplicated? (1: present, 0: absent)