doi:10.1038/nature09923

**Sections**



## S1. Data and Methods

We used lexical data to construct posterior probability distributions of the language phylogenies for the Austronesian, Bantu, Indo-European and Uto-Aztecan language families. The structural data was mapped onto these phylogenies using a continuous-time Markov model of trait evolution under dependent and independent models of trait evolution. The relative fit of these dependent and independent models was calculated across the posterior tree distribution for each comparison in each family.

### S1.1. Lexical Data

The language data used to build the phylogenetic trees in the posterior tree distribution were derived from four large datasets of basic vocabulary, covering the Austronesian, Bantu, Indo-European and Uto-Aztecan families. Basic vocabulary (e.g. words for body parts, kinship terms, simple verbs etc) was used as these items of lexicon are resistant to being borrowed between languages and stable over time (Embleton 1986, Greenhill et al. 2008, 2009), and are closely linked to population history (Mace and Pagel 1994) in both Europe and the Pacific (Gray & Atkinson 2003, Gray et al. 2009, Greenhill et al. 2008, Hunley et al. 2008, Lansing et al. 2007).

In each of the four language families the homologous (cognate) items of basic vocabulary were identified using systematic sound correspondences following the linguistic comparative method (e.g. Durie and Ross 1993). We encoded these sets of cognate words into binary characters representing the presence or absence of each cognate set in each language. This process is described more fully elsewhere (Gray & Atkinson 2003, Gray et al. 2009).

The Austronesian lexical data were obtained from the Austronesian Basic Vocabulary Database (Greenhill et al. 2008, http://language.psy.auckland.ac.nz). From this database we extracted the cognate sets for the 210-item wordlist for 400 languages. We used two non-Austronesian languages to "root" the trees - an archaic variant of the Sino-Tibetan language Chinese that was spoken between 2,300 and 2,900 years ago, and the Tai-Kadai language Buyang (Gray et al. 2009). Cognate judgments were made and checked by historical linguists with expertise in the region.

The Bantu phylogenetic trees were constructed from the lexical data set previously used by Holden (2002). This dataset was derived from Bastin et al (1999) and contains cognate information from a 92 item basic vocabulary wordlist. The trees contained 73 Bantu languages and 2 Bantoid languages. The Bantoid languages, Tiv and Ejagham, were used to root the trees.

The Indo-European lexical data based on the published dataset of Dyen et al. (1992). We used a substantially expanded version of the data described in Gray and Atkinson (2003) with 82 distinct languages (including 8 new to this study) and 4049 cognate sets. Cognate identifications were validated against etymological dictionaries and borrowings were removed. In these analyses the trees were rooted in the language Hittite.

The Uto-Aztecan phylogenetic trees were produced from lexical data coded by Ross et al. (in prep.). This dataset contains cognate information for 34 languages from a 108 item basic vocabulary list.

### S1.2. Language Phylogenies

Posterior tree distributions for all four lexical datasets were estimated using a Bayesian Markov Chain Monte Carlo approach implemented in the program *BayesPhylogenies* (Pagel and Meade 2004, 2005). This Bayesian approach samples from the posterior probability distribution the trees that explain the data well under a given model of evolution. Following Gray et al (2009), we used a two-state covarion model of lexical evolution. The covarion model allows cognate sets to change between different rates at different points along the tree. We ran the Bantu, Indo-European, and Utoaztecan data for five million generations as plots of the critical parameters (log likelihood, tree length) showed that this was sufficient for convergence. To avoid auto-correlation we sampled trees from the posterior every 5,000 generations. Burn-in was set to two million generations after inspection of the log-likelihood plots and these initial trees were discarded leaving a total of 600 trees in the posterior sample.

Due to the much larger size of the Austronesian data this analysis was run for much longer – 100 million generations over six replicates – to ensure adequate mixing (Gray et al. 2009). Trees were sampled every 100,000 generations and burn-in was set to 40 million generations after inspection of critical parameters showed that convergence had been reached. This left a total of 4,200 trees in the posterior tree sample.

### S1.3. Structural Data

Greenberg's word order universals are an empirically determined set of co-occuring traits. Two of these are reproduced below:

a. 'Languages with dominant VSO (Verb-Subject-Object) order are always prepositional' (Greenberg's Universal 3)
b. 'With overwhelmingly greater than chance frequency, languages with normal SOV (Subject-Object-Verb) order are postpositional' (postpositions are just like prepositions, except that they follow the noun rather than precede it). (Greenberg's Universal 4)

Thus for sentences corresponding to the English 'The man put the dog in the canoe', (a) predicts that languages which express the action using the order of elements 'Put man dog' will express the location with an element ordered 'in canoe', while (b) predicts that languages which use 'Man dog put' will express the location as 'canoe in'. Greenberg also made many further predictions of correlated word-orders, e.g. (Universal 2) that genitive (possessive) nouns tend to precede their head nouns in languages with prepositions ('the prime minister of New Zealand), and tend to follow them in languages with postpositions ('New Zealand's prime minister').

Greenberg's observations have been reformulated by Dryer on the basis of extensive testing. The typological data and coding principles used in this study are derived from Dryer's word order typology data published in Haspelmath et. al (2005). A form of this data is also available online at http://wals.info (Dryer 2008a-h). The Austronesian and Indo-European databases were supplemented with material from published materials (see S2.1, S2.3). The Bantu family is relative young, and for some of the features under consideration the Bantu languages show little or no variation. Where there is no variation it is not meaningful to calculate statistical measures of correlation, and so these features have been excluded from the analysis. The features affected are: Adposition~Noun, Genitive~Noun, Relative clause~Noun, and Object~Verb. For the purpose of the analysis the feature states have been recoded as binary, with possible polymorphisms (recoded as "01"). Only states which are attested in the four language families under consideration are recoded (so not, for example, DRYADP state 3). There are a few instances of states which cannot be coded, these are coded as missing: "-". The mappings between the WALS coding and the binary coding used in the present study are as follows:

**Adjective-Noun order (DRYADJ)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Modifying adjective precedes noun (AdjN) |
| 2 | 1 | Modifying adjective follows noun (NAdj) |
| 3 | 01 | Both orders of noun and modifying adjective occur, with neither dominant |

**Adposition-Noun phrase order (DRYADP)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Postpositions |
| 2 | 1 | Prepositions |
| 4 | 01 | More than one adposition type with none dominant |

State 3 did not occur in Austronesian or Indo-European languages.

**Demonstrative-Noun order (DRYDEM)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Demonstrative word precedes noun (DemN) |
| 2 | 1 | Demonstrative word follows noun (NDem) |
| 4 | 1 | Demonstrative suffix on noun |
| 5 | 01 | Demonstrative simultaneously before and after noun |
| 6 | - | Two or more of above types with none dominant |

State 6 from the original data can't be treated as polymorphism since "two of the above" might mean states 2 and 4.

**Genitive-Noun order (DRYGEN)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Genitive-noun (GenN) |
| 2 | 1 | Noun-nenitive (NGen) |
| 3 | 01 | Both orders occur with neither order dominant |

**Numeral-Noun order (DRYNUM)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Numeral precedes noun (NumN) |
| 2 | 1 | Numeral precedes noun (NumN) |
| 3 | 01 | Both orders of numeral and noun with neither order dominant |

**Object-Verb order (DRYOBV)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Object precedes verb (OV) |
| 2 | 1 | Object follows verb (VO) |
| 3 | 01 | Both orders with neither order dominant |

**S2.1.7.Relative clause-Noun order (DRYREL)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Relative clause follows noun (NRel) |
| 2 | 1 | Relative clause precedes noun (RelN) |
| 4 | - | Correlative relative clause |
| 7 | 01 | Mixed types of relative clause with none dominant |

**Subject-Verb order (DRYSBV)**

| *original* | *recoded* | *WALS coding information* |
|---|---|---|
| 1 | 0 | Subject precedes verb (SV) |
| 2 | 1 | Subject follows verb (VS) |
| 3 | 01 | Both orders with neither order dominant |

Dryer's criteria for classifying languages as having one or another type of dominant word order are set out in Dryer 2005 a, b, c (330-331; 338-339; 371). Note that Dryer's definitions of 'dominant word order' conflict in some cases with notions of 'basic word order' appearing in the literature. Basic word order, which is not considered by Dryer, is the underlying order required by a particular theoretical analysis, but which is not necessarily reflected in surface structure or discourse frequency (see e.g. Kayne 1994). We agree with Dryer that 'basic word order' is too theoretically loaded a concept to be appropriate for use in an empirical study in language variation.

Readers may note that some languages with flexible word order are coded as having a dominant word order; most of the Slavonic languages, such as Russian or Polish, have considerable flexibility in their word order, but apart from Lusation and Belorussian, Dryer has these classified as having dominant VO due to higher frequency in texts and pragmatic neutrality. In addition, some languages are coded as having no dominant word order (i.e. polymorphic). In the original coding these languages are Belorussian, Dutch, German and Dutch/Flemish. Dryer's reasons for coding these languages as having no dominant order of object and verb are set out explicitly:

> Languages in which neither OV nor VO is dominant fall into two sorts. [… The] second class of language in which both OV and VO are common are languages in which word order is primarily determined syntactically, but in which there are competing OV and VO constructions. German is an instance of this, in that VO order is used in main clauses in which there is no auxiliary verb […], while OV order is used in clauses with an auxiliary verb […], and in subordinate clauses introduced by a subordinator […]. (Dryer 2005:338)

Polymorphism in dominant order of object and verb is thus a linguistic reality, and one which is an inherited feature of the Germanic languages (Davis 2002).

Two languages in the database have adposition order coded as polymorphic: Wakhi and Waziri/Pashto (both Indo-European). Prepositions and postpositions involve different lexemes in different constructions, and there is never free choice about whether a particular adposition is used as one type or the other. The case for coding these languages as polymorphic is thus very strong.

## S1.4. Correlation Analysis

To identify the relationships between the word order characters we mapped the WALS typological data onto these phylogenies using a continuous-time Markov model of trait evolution under a dependent and independent model of trait evolution. The linguistic characters were adapted from Dryer's word order features, published in the WALS database (Dryer 2005a-h; Section S2). The phylogenetic correlation analysis was carried out using DISCRETE, a Bayesian method of model comparison for discrete parameters implemented in *BayesTraits* (Pagel and Meade 2006). DISCRETE*/BayesTraits* estimates the posterior probability distribution of transition rates between the character states across the set of phylogenetic trees (Section S1.1) under a continuous-time Markov model of trait evolution (Pagel 1994, Pagel and Meade 2006).

We calculated the fit of an independent and dependent model of trait evolution on each pair of characters. In the independent model, the linguistic features evolve separately (Figure S1). For example, a transition from having postpositions to having prepositions is not contingent on the background state of the second character (word order) and vice versa. Under this model there are four rates of change between states: Postpositions ⇒ Prepositions, Prepositions ⇒ Postpositions, Object-Verb ⇒ Verb-Object and Verb-Object ⇒ Object-Verb.

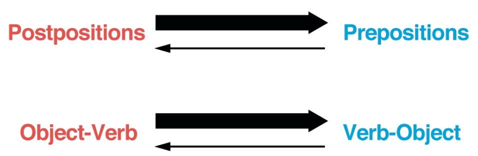


*Figure S1: An independent model of linguistic trait evolution. Changes in character A (Adposition type, top) are not linked to changes in character B (word order, bottom). The magnitude of the estimated transition parameters is indicated here by arrow weight.*

In contrast, the dependent model treats each pair of linguistic features as a single, four-state character, which can be represented by model with eight rates of change (Figure S2).

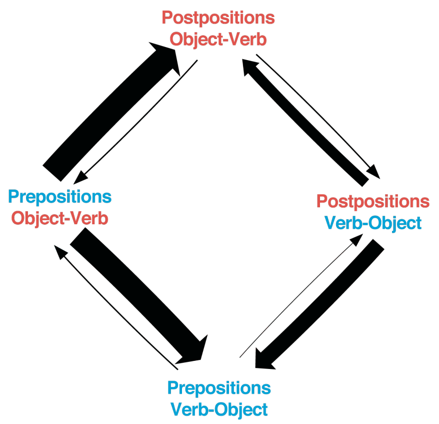


*Figure S2: A dependent model of linguistic trait evolution. Changes in character A (Adposition type) are linked to changes in character B (word order). The magnitude of the estimated transition parameters is indicated here by arrow weight.*

The fit of these two models can be compared using the likelihood ratio test (Edwards 1972),

$$BF = -2\log(L(H_{dep})/L(H_{ind}))$$

A positive Bayes Factor indicates that the 'dependent' hypothesis ($H_{dep}$) is preferred over the independent one ($H_{ind}$). A Bayes factor of 5 is considered strong evidence, and 10 is extremely strong. In the examples of dependent and independent models given above, the dependent model clearly captures the behavior of a system in which the states 'VO + prepositions' and 'OV + postpositions' are much more likely than the states 'VO + postpositions', 'OV + prepositions'. The highest likelihood parameters of the independent model only capture the e.g. high frequency of 'VO' and 'prepositions' in the Austronesian data set.

Note that the independent model can be expressed as a special case of the dependent model, so the test cannot show that the independent model is preferred over the dependent; rather it can only show that the dependent model is not superior.

It was observed that the Markov chain occasionally visited a very low likelihood tree, leading to occasional downwards "steps" in the trace of the harmonic mean of log likelihood scores. This tended to make the Bayes factor measurement unreliable for short (e.g. 10 million generations). Ultimately, six runs of 1,000,000,000 iterations were carried out for each condition (dependent and independent tests for each feature pair for each lineage) with a 5,000,000 iteration burn-in, sampling every 10,000 iterations. The results of each billion-iteration run were inspected to check stability, and four additional runs were carried out in all cases where there was any instability around the Bayes Factor value of 5, the cutoff for a 'strong' dependency. The reported result for each condition is the median value from all runs (i.e. 6 or 10 billion iterations).

## S2. Language data

The World Atlas of Linguistic Structures lacks data from several languages near the roots of the Indo-European and Austronesian phylogenetic trees. The phylogenetic comparative method performs best with deep and highly resolved phylogenetic trees, and lacking these basal languages the statistical power of the method is considerably reduced. To increase the power of the analyses additional taxa were coded.

### S2.1. Austronesian Data

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Agta | agt | agc | 0 | 1 | 1 | 1 | 0 | 1 | - | 1 | |
| Alune | alp | aln | 1 | 1 | 1 | 0 | 1 | 1 | - | 0 | |
| Ambai | amk | amq | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | |
| Anejom | aty | anj | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | |
| Aria | mwh | ama | 1 | 1 | 1 | - | 1 | 1 | - | 0 | |
| Babatana | baa | sis | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Bajo | bdl | baj | 1 | 1 | 1 | 1 | - | 1 | 0 | 01 | |
| Balangaw | blw | blg | - | - | 0 | - | - | 1 | - | - | |
| Banoni | bcm | bnn | 1 | 1 | 0 | 1 | 1 | 1 | - | 0 | |
| Bikol | bcl | bkl | 01 | 1 | - | 01 | - | 1 | - | 1 | |
| Bontok (Guinaang) | bnc | btk | 01 | 1 | - | 1 | 0 | 1 | 0 | 01 | |

doi:10.1038/nature09923


| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Buginese (Soppeng Dialect) | bug | bug | - | - | 1 | 1 | - | 1 | 0 | - | |
| Buma | tkw | bum | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Buru | mhs | buu | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | |
| Bwaidoga | bwd | idn | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| Cebuano | ceb | ceb | - | 1 | - | 1 | - | 1 | 0 | 1 | |
| Central Amis | ami | ami | - | - | - | - | - | 1 | 1 | 1 | Chen, Teresa M. 1987. *Verbal constructions and verbal classification in Nataoran-Amis*. Pacific Linguistics C-85. Canberra: Pacific Linguistics. |
| Chamorro | cha | cha | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |
| Ci'uli Atayal | tay | ata | 01 | 1 | 1 | 1 | 0 | 1 | - | 1 | Holmer, Arthur. 1993. Atayal clitics and sentence structure. *Working papers Lund University Department of Linguistics*. 71-95. |
| Dehu | dhv | dre | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Dobuan | dob | gmw | 1 | 0 | 1 | 0 | 1 | 0 | 01 | 0 | |
| East Sumbanese (Southern Kambera Dialect) | xbr | kam | 1 | 1 | - | - | - | 1 | 0 | 1 | |
| East Sumbanese (Lewa Dialect) | xbr | kam | 1 | 1 | - | - | - | 1 | 0 | 1 | |
| East Sumbanese (Umbu Ratu Nggai Dialect) | xbr | kam | 1 | 1 | - | - | - | 1 | 0 | 1 | |
| Fijian (Bau) | fij | fij | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Futuna | fut | fut | 1 | 1 | 1 | 1 | 01 | 1 | 0 | 0 | |
| Futuna (East) | fud | fut | 1 | 1 | 1 | 1 | 01 | 1 | 0 | 0 | |
| Futuna (West) | fut | fut | 1 | 1 | 1 | 1 | 01 | 1 | 0 | 0 | |
| Gapapaiwa | pwg | gap | - | 0 | - | 0 | 1 | 0 | 0 | 0 | |
| Gimán | gzn | tab | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | |
| Hawaiian | haw | haw | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |
| Iaai | iai | iaa | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Iban | iba | iba | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Ifugao (Batad) | ifb | ifu | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |
| Ilokano | ilo | ilo | 0 | 1 | 0 | - | - | 1 | - | 1 | |
| Indonesian | ind | ind | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Kairiru | kxa | krr | 1 | 01 | 1 | 1 | 1 | 0 | 0 | 0 | |
| Kambera | xbr | kam | 1 | 1 | - | - | - | 1 | 0 | 1 | |
| Kapampangan | pam | kpm | 01 | 1 | - | 1 | 0 | 1 | 0 | 1 | |
| Kasira | irh | irr | 1 | 1 | 1 | 0 | 1 | 1 | - | 0 | |
| Kilivila | kij | klv | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | |
| Kiribati | gil | krb | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Kokota | kkk | kkt | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |
| Kove | kvc | kkv | 1 | 1 | 1 | 0 | 1 | 1 | - | 0 | |
| Kuanua | ksd | tla | 01 | 1 | 0 | 1 | 0 | 1 | 0 | 01 | |
| Kusaie | kos | kos | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Kwaio | kwd | kwa | 1 | 1 | 1 | - | 0 | 1 | - | 0 | |
| Lakalai | nak | nak | 1 | 1 | 1 | 1 | 01 | 1 | - | 0 | |
| Lamaholot | slp | lmh | - | 01 | 1 | 0 | 1 | 1 | - | 0 | |
| Lampung | ljp | lmp | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Lenakel | tnl | len | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Letinese | lti | let | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | |
| Loniu | los | lon | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Lou | loj | lou | 1 | 1 | 1 | 1 | - | 1 | - | 0 | |
| Ma’anyan | mhy | myn | 1 | - | 1 | 1 | 1 | 1 | 0 | 0 | |
| Malay (Bahasa Indonesia) | ind | ind | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Maleu | mgl | mlu | 1 | 1 | 1 | 01 | 1 | 1 | 0 | 0 | |
| Mamanwa | mmn | mmn | 01 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |
| Manam | mva | mnm | 1 | 0 | - | 1 | 1 | 0 | 0 | 0 | |
| Manggarai | mqy | mng | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Maori | mri | mao | 1 | 1 | - | 1 | 1 | 1 | 0 | 1 | |
| Maututu | nak | nak | 1 | 1 | 1 | 1 | 01 | 1 | - | 0 | |
| Mekeo | mek | mke | 1 | 0 | - | 0 | - | 0 | - | 0 | |
| Merina | plt | mal | 1 | 1 | 01 | 1 | 1 | 1 | 0 | 1 | |
| Minangkabau | min | min | 1 | 1 | 1 | 1 | 0 | 01 | 0 | 01 | |
| Mokilese | mkj | mok | 1 | 1 | 1 | 01 | 1 | 1 | 0 | 01 | |
| Mor | mhz | mor | 1 | 1 | 1 | 0 | 1 | 1 | - | 0 | |
| Motu | meu | mtu | 1 | 0 | - | 0 | - | 0 | - | 0 | |
| Mouk | mwh | ama | 1 | 1 | 1 | - | 1 | 1 | - | 0 | |
| Muna (Katobu-Tongkuno Dialect) | mnb | mna | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Murnaten | alp | | 1 | 1 | 1 | 0 | 1 | 1 | - | 0 | |
| Mussau | emi | mus | 0 | - | 1 | 1 | 0 | 1 | 0 | 0 | |
| Nakanai (Bileki Dialect) | nak | nak | 1 | 1 | 1 | 1 | 01 | 1 | - | 0 | |
| Nalik | nal | nal | 1 | 1 | 1 | - | 1 | 1 | 0 | 0 | |
| Nehan Hape | nsn | neh | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Nengone | nen | nne | - | - | 1 | 1 | 0 | 1 | - | 01 | |
| Ngadha | nxg | ngd | 1 | 1 | - | 1 | 01 | 1 | - | 0 | |
| Nggela | nlg | gel | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Nguna | llp | ngu | 1 | 1 | 1 | 1 | 1 | 1 | - | 0 | |
| Nias | nia | nia | 1 | 1 | 1 | 1 | 01 | 1 | 0 | 1 | |
| Nissan | nsn | neh | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Niue | niu | niu | 1 | 1 | - | 1 | 0 | 1 | 0 | 1 | |
| Paamese South | pma | pms | 1 | 1 | - | 1 | 1 | 1 | 0 | 0 | |
| Paiwan | pwn | pai | 01 | 1 | 1 | 1 | 01 | 1 | 01 | 1 | Chang, Anna Hsiou-chuan. 2006. A Reference Grammar of Paiwan. PhD Thesis, Australian National University. |
| Palauan | pau | pal | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | |
| Pangasinan | pag | pnn | 01 | 1 | - | - | 0 | 1 | - | 1 | |
| Paulohi | plh | plh | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | |
| Phan Rang Cham | cjm | cme | 1 | - | 1 | 1 | - | 1 | - | 0 | |
| Ponapean | pon | poh | 1 | 1 | 1 | 01 | 1 | 1 | 0 | 0 | |
| Popalia | bhq | tuk | 1 | 1 | 1 | 1 | 1 | 1 | 01 | 1 | |
| Pulo-Annan | sov | son | 01 | 1 | 1 | 1 | 0 | 1 | - | 0 | |
| Puluwatese | puw | pul | - | 1 | 1 | 0 | 0 | 1 | 0 | 0 | |
| Rapanui | rap | rap | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Rotuman | rtm | rot | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Roviana | rug | rov | 1 | - | 1 | 1 | 0 | 1 | 0 | 1 | |
| Rukai | dru | ruk | 0 | 1 | 0 | 1 | 0 | 1 | - | 1 | Li, Paul J. 1996. The prominal systems in Rukai. In Nothofer, Bernd (editor). *Reconstruction, classification, description: Festschrift in honor of Isidore Dyen*. Hamburg: Abera Verlag. 209-230. |
| Sa'a | apb | saa | 1 | 1 | - | 1 | 0 | - | - | - | |
| Sakao | sku | sak | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Saliba | sbe | slb | 1 | 0 | 0 | 0 | 1 | 0 | 0 | - | |
| Samoan | smo | sam | 1 | 1 | - | 1 | 1 | 1 | 0 | 1 | |
| Sediq | trv | see | 01 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | Tsukida, Naomi. 2005. Seediq. In Adelaar, Alexander and Nikolaus P. Himmelmann (editors). The Austronesian languages of Asia and Madagascar. London and New York: Routledge. 291-325. |
| Sengseng | ssz | kau | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | |
| Siar | sjr | sir | 0 | 1 | - | 1 | 0 | 1 | 0 | 0 | |
| Siraya | fos | | - | 1 | 1 | 1 | 0 | 1 | 0 | 1 | Adelaar, Alexander. 1997. Grammar Notes on Siraya, an Extinct Formosan Language. Oceanic Lingusitics. 26:2, 362-397. Tsuchida, Shigeru. 1996. Personal pronouns in Siraya (Formosa) In Nothofer, Bernd (editor). *Reconstruction, classification, description: Festschrift in honor of Isidore Dyen*. Hamburg: Abera Verlag. 231-248 |
| Sisingga | baa | sis | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Squliq Atayal | tay | ata | 01 | 1 | 1 | 1 | 0 | 1 | - | 1 | Holmer, Arthur. 1993. Atayal clitics and sentence structure. *Working papers Lund University Department of Linguistics*. 71-95. |
| Sunda | sun | sun | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Sye | erg | err | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Tagalog | tgl | tag | 01 | - | - | 1 | 0 | 1 | 0 | 1 | |
| Tahitian | tah | tah | 1 | 1 | 0 | 01 | 01 | 1 | 0 | 1 | |
| Tai'of | sps | taf | 1 | - | 1 | 1 | 0 | 1 | 0 | 0 | |
| Tboli | tbl | tbo | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Teanu | tkw | bum | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Teop | tio | teo | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Tigak | tgc | tgk | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Timugon | tih | tim | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Toba Batak | bbc | bto | 1 | 1 | 1 | 1 | 01 | 1 | 0 | 1 | |
| Tokelau | tkl | tke | - | 1 | 1 | - | - | 1 | 0 | 1 | |
| Tongan | ton | tng | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | |
| Tsou | tsu | | 0 | 1 | 1 | 1 | 0 | 1 | 01 | 1 | Zeitoun, Elizabeth. 2005. Tsou. In Adelaar, Alexander and Nikolaus P. Himmelmann (editors). The Austronesian languages of Asia and Madagascar. London and New York: Routledge. 259-290. |
| Tugun | tzn | tgn | 1 | - | 1 | 0 | 1 | 1 | 0 | 0 | |
| Tungak | lcm | tnk | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Ura | uur | ura | 1 | - | - | 1 | 1 | 1 | - | 0 | |
| Vitu | wiv | bvi | 1 | 1 | 1 | 1 | - | 1 | 0 | 0 | |
| Western Bukidnon Manobo | mbb | mwb | 0 | 1 | 0 | 1 | - | 1 | - | 1 | |
| Woleai | woe | bum | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |
| Woleaian | woe | wol | 1 | 1 | 1 | 01 | 0 | 1 | 0 | 0 | |
| Wolio | wlo | wlo | 1 | 1 | 0 | - | 0 | 1 | 0 | 1 | |
| Wuna | mnb | mna | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |

## S2.2 Bantu data

| NAME (Guthrie Code) | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV |
|---|---|---|---|---|---|---|---|---|---|---|
| Bemba (M42) | bmy | bem | 1 | - | - | - | 1 | - | - | - |
| Bira (D322) | brf | bia | 1 | 1 | - | 1 | - | 1 | - | 0 |
| Bubi (A31) | bvb | bub | 1 | - | 0 | - | - | 1 | - | 0 |
| Caga (E621) | old | cga | 1 | - | - | - | - | 1 | - | 0 |
| Cewa (N31b) | nya | cic | 1 | 1 | 1 | 1 | 1 | 1 | - | 0 |
| Doko (C40D) | ngc | dok | - | 1 | 1 | - | - | 1 | - | 0 |
| Duala (A242) | dua | dua | 1 | - | 0 | 1 | - | 1 | - | 0 |
| Ganda (J15) | lug | lda | 1 | 1 | 1 | 1 | 1 | - | 0 | 0 |
| Gikuyu (E51) | kik | kik | 1 | 1 | 1 | 1 | - | 1 | 0 | 0 |
| Hadimu (G43c) | swh | swa | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Haya (J22H) | hay | hya | 1 | 1 | 1 | - | 1 | 1 | 0 | 0 |
| Hima (J13) | nyn | rny | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Hunde (J51) | hke | hnd | 1 | - | 1 | 1 | 1 | 1 | 0 | 0 |
| Kaguru (G12) | kki | kgr | 1 | 1 | 1 | - | 1 | 1 | 0 | - |
| Kamba (E55) | kam | kba | 1 | - | 1 | 1 | - | 1 | - | 0 |
| Lamba (M54) | lam | lmb | 1 | 1 | - | 1 | - | - | - | - |
| Lega (D253) | lgm | leg | - | - | - | - | - | 1 | 0 | 0 |

| NAME (Guthrie Code) | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV |
|---|---|---|---|---|---|---|---|---|---|---|
| Lingala (C36) | lin | lin | 1 | - | - | 1 | - | 1 | - | 0 |
| Luba (L33) | lua | blu | 1 | 1 | - | - | - | 1 | - | - |
| Lwena (K141) | lue | luv | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Makwa (P311) | vmw | mua | 1 | 1 | - | - | 1 | 1 | - | 0 |
| Mambwe (M15) | mgr | mmw | 1 | - | - | - | 1 | - | - | - |
| MongoNkundo (C611) | lol | mgo | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Mongo (C613) | lol | mgo | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Mpongwe (B11a) | mye | mpo | 1 | - | 1 | - | 1 | 1 | - | 01 |
| Ndebele (S44) | nbl | ndb | 1 | - | 1 | 1 | - | 1 | - | 0 |
| Ndembu (K221) | lun | lbu | 1 | 1 | - | 1 | 1 | 1 | 0 | 0 |
| Ndonga (R22) | ndo | ndo | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Ngombe (C41) | ngc | nbe | - | - | - | - | - | 1 | - | 0 |
| Ngoni (S45) | nya | cic | 1 | 1 | 1 | 1 | 1 | 1 | - | 0 |
| Nyamwezi (F22) | nym | nym | 1 | - | 1 | 1 | 1 | 1 | - | 01 |
| Nyanja (N31a) | nya | cic | 1 | 1 | 1 | 1 | 1 | 1 | - | 0 |
| Rundi (J62) | run | rnd | - | - | - | - | - | - | - | 0 |
| Rwanda (J611) | kin | kin | 1 | 1 | 0 | - | 1 | 1 | - | 0 |
| Sena (N44) | seh | sen | 1 | 1 | - | - | - | - | - | - |
| Shambala (G23) | ksb | shm | 1 | - | 0 | - | - | 1 | - | 0 |
| Shona (S10) | sna | shn | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Songe (D10S) | sop | sge | - | 1 | - | - | - | - | - | - |
| Sotho (S33) | nso | stn | - | - | - | 1 | - | 1 | - | 0 |
| Sukuma (F21) | suk | skm | - | - | - | - | - | 1 | - | 0 |
| Swati (S43) | ssw | swt | 01 | - | 0 | 1 | 01 | 1 | 0 | 01 |
| Tiv (802) | tiv | tiv | 01 | 1 | 1 | - | - | 1 | - | 0 |
| Tonga (M641) | toi | toz | 1 | - | 1 | 1 | 1 | - | 0 | - |
| Tsogo (B31) | tsv | tgo | 1 | 1 | 1 | - | 1 | - | - | - |
| Venda (S21) | ven | ven | 1 | - | - | - | 1 | 1 | - | 0 |
| Xhosa (S41) | xho | xho | 1 | 1 | - | 1 | 1 | 1 | 0 | 0 |
| Yao (P212) | yao | yao | 1 | - | 1 | - | - | 1 | - | 01 |
| Zulu (S42) | zul | zul | 1 | 1 | - | 1 | - | 1 | 0 | 0 |

## S2.3 Indo-European data

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghan | prs | | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | Windfuhr, Gernot and John R. Perry. 2009. Persian and Tajik. In: Windfuhr, Gernot. The Iranian Languages. London/New York: Routledge. 416-544. |
| Afrikaans | afr | | 0 | 1 | 0 | 0 | 0 | 01 | 0 | 0 | Donaldson, Bruce. 1994. Afrikaans. In: König, Ekkehard and Johan van der Auwera. The Germanic Languages. London/New York: Routledge. 478-504; Donaldson, Bruce. 1993. A Grammar of Afrikaans. Berlin/New York: Mouton de Gruyter. |

doi:10.1038/nature09923 

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Albanian G | als | alb | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | |
| Ancient Greek | | | 01 | 0 | 0 | 1 | 0 | 01 | 0 | 0 | Luraghi, Silvia, Anna Pompei and Stavros Skopeteas. 2005. Ancient Greek. München: Lincom. |
| Armenian Mod | hye | arw | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Assamese | asm | ass | 0 | 0 | 0 | 0 | 01 | 0 | - | 0 | Goswami, G.C. and Jyotiprakash Tamuli. 2003. Asamiya. In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 391-443. |
| Baluchi | bgp | | 0 | 01 | 0 | 0 | 0 | 0 | 0 | 0 | Jahani, Carina and Agnes Korn. 2009. Balochi. In: Windfuhr, Gernot. The Iranian Languages. London/New York: Routledge. 634-692. |
| Bengali | ben | ben | 0 | 0 | 0 | - | 0 | 0 | - | 0 | Dasgupta, Probal. 2003. Bangla. In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 351-390. |
| Bihari | mai | mai | 0 | 0 | 0 | - | 0 | 0 | 01 | 0 | Davis, Alice I. 1984. Basic colloquial Maithili. Dehli. Motilal Banarsidass; Yadav, Ramawatar.1996. A reference grammar of Maithili. Berlin/New York: Mouton de Gruyter. |
| Breton ST | bre | bre | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Bulgarian | bul | bul | 0 | 1 | 0 | 01 | 0 | 1 | 0 | 01 | |
| Byelorussian | bel | blr | 0 | 1 | 0 | 1 | 0 | 01 | 0 | 01 | Sussex, Roland. 2006. The Slavic Languages. Cambridge. CUP; Mayo, Peter. 1993. Belorussian. In: Comrie, Bernard and Greville G. Corbett. The Slavonic Languages. London/New York: Routledge. 887-946. |
| Catalan | cat | ctl | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 01 | |
| Cornish | cor | crn | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | |
| Czech | ces | cze | 0 | 1 | 0 | 01 | 0 | 1 | 0 | 0 | |
| Danish | dan | dsh | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | |
| Dutch List | nld | dut | 0 | 1 | 0 | 1 | 0 | 01 | 0 | 0 | |
| English ST | eng | eng | 0 | 1 | 0 | 01 | 0 | 1 | 0 | 0 | |
| Faroese | fao | | 0 | 1 | 0 | 0 | 0 | 1 | - | 0 | Barnes, Michael P. and Eivind Weyhe.1994. Faroese. In: König, Ekkehard and Johan van der Auwera. The Germanic Languages. London/New York: Routledge. 190-219. |
| Flemish | zea | | 0 | 1 | 0 | 1 | 0 | 01 | 0 | 0 | |
| French | fra | fre | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | |
| Frisian | frs | fri | 0 | 1 | 0 | 01 | 0 | 01 | 0 | 0 | Hoekstra, Jarich and Peter Meijes Tiersma. Frisian. In: König, Ekkehard and Johan van der Auwera. The Germanic Languages. London/New York: Routledge. 505-531. |
| German ST | deu | ger | 0 | 1 | 0 | 1 | 0 | 01 | 0 | 0 | |
| Gothic | got | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | Wright, Joseph and O. L Sayce. 1954. Grammar of the Gothic Language. Oxford: Clarendon Press; Lehmann, Winfred. 1994. Gothic and the reconstruction of proto-Germanic. In: König, Ekkehard and Johan van der Auwera. The Germanic Languages. London/New York: Routledge. 19-37; Ferraresi, Gisella. 2005. Word order and phrase structure in Gothic. Leuven: Peeters Pub & Booksellers; Krause, Wolfgang. 1953. Handbuch des Gotischen. München: C. H. Beck. |
| Greek Mod | ell | grk | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 01 | |
| Gujarati | guj | guj | 0 | 0 | 0 | - | 0 | 0 | - | 0 | |
| Hindi | hin | hin | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | |
| Hittite | hit | | 0 | 0 | 0 | 0 | 0 | 0 | 01 | 0 | Luraghi, Silvia. 1990. Old Hittite Sentence Structure. London/ New York: Routledge; Luraghi, Silvia. 1997. Hittite. München: Lincom; Luraghi, Silvia. 1998. The Anatolian Languages. In: Ramat, Anna and Paolo Ramat. The Indo-European Languages. London/New York: Routledge. 169-196; Sturtevant, E. H. 1933. A Comparative Grammar of the Hittite Language. Yale: University Press. |
| Icelandic ST | isl | ice | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | |

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Irish B | gle | iri | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Italian | ita | ita | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 01 | |
| Kashmiri | kas | kas | 0 | 0 | 0 | 0 | 0 | 1 | 01 | 0 | |
| Kurdish | kmr | krd | 1 | 01 | 01 | 1 | 0 | 0 | 0 | 0 | McCarus, Ernest N. 2009. In: Windfuhr, Gernot. The Iranian Languages. London/New York: Routledge. 587-633; Soane, E. B. 1913. Grammar of the Kurmanji or Kurdish Language. London: Luzac and Co. |
| Ladin | lld | | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | Haiman, John. 1988. Rhaeto-Romance. In: Harris, Martin and Nigel Vincent. The Romance Languages. London/Sydney: Coom Helm. 351-390; http://din.micura.it/voc_vb/lad/index.html; Valentini, Erwin. 2001. Gramatica dl Ladin Standard. Union Generela di Ladins dles Dolomites. |
| Lahnda | prb | | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | Shackle, Christopher. 2003. Panjabi. In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 581-621. |
| Latin | | | 01 | 1 | 01 | 01 | 0 | 01 | 0 | 01 | Vincent, Nigel. 1988. Latin. In: Harris, Martin and Nigel Vincent. The Romance Languages. London/Sydney: Coom Helm. 26-78; Peranteau, Paul; Judith Levi and Gloria Phares. 1972. The Chicago which hunt: papers from the Relative Clause Festival, April 13, 1972. Chicago: CUP; Bauer, Brigitte. 1995. The Emergence and Development of SVO patterning in Latin and French. New York: Oxford US. |
| Latvian | lav | lat | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 01 | |
| Lithuanian ST | lit | lit | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | |
| Lusatian L | hsb | srb | 0 | 1 | 0 | 01 | 0 | 0 | 0 | 0 | Stone, Gerald. 1993. Sorbian. In: Comrie, Bernard and Greville G. Corbett. The Slavonic Languages. London/New York: Routledge. 593–685; Berger, Tilman. 2007. Tendencies of the evolution of a definite article in the Sorbian languages. |
| Lusatian U | hsb | | 0 | 1 | 0 | 01 | 0 | 0 | 0 | 0 | Stone, Gerald. 1993. Sorbian. In: Comrie, Bernard and Greville G. Corbett. The Slavonic Languages. London/New York: Routledge. 593–685; Schaarschmidt, Gunter. 2002. Upper Sorbian. München: Lincom. |
| Luxembourgish | ltz | | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | Russ, Charles V. J. 1996. Lëtzebuergesch: A Linguistic Description. In: Newton, Gerald. Luxembourg and Lëtzebuergesch. Language and Communication at the Crossroads of Europe. Oxford: Claredon Press. 67-95. |
| Macedonian | mkd | mcd | 0 | 1 | 0 | 01 | 0 | 1 | 0 | 0 | |
| Marathi | mar | mhi | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| Marwari | rwr | skw | 0 | 0 | 0 | 0 | 0 | 0 | 01 | 0 | Gusain, Lakhan. 2004. Marwari. München: Lincom. |
| Nepali List | nep | nep | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | Riccardi, Theodore. 2003. Nepali. In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 538-580. |
| Old Church Slavonic | chu | | 0 | 1 | - | 01 | 0 | 01 | 0 | 01 | Huntley, David. 1993. Old Church Slavonic. In: Comrie, Bernard and Greville G. Corbett. The Slavonic Languages. London/New York: Routledge. 125-187; Lunt, Horace G. 1965. Old Church Slavonic Grammar. 's-Gravenhage: Mouton de Gruyter. |
| Old English | ang | | 0 | 1 | 0 | 01 | 0 | 01 | 0 | 0 | Mitchell, Bruce and Fred Robinson. 1992 [1964]. A Guide to Old English. Malden/Oxford: Blackwell; Lass, Roger. 1995. Old English. A Historical Companion. Cambridge: CUP; Fischer, Olga. 2000. The Syntax of Early English. Cambridge: CUP. |
| Old Norse | non | | 0 | 1 | 0 | 01 | 0 | 1 | 0 | 0 | Faarlund, Jan T. 1994. Old and Middle Scandinavian. In: König, Ekkehard and Johan van der Auwera. The Germanic Languages. London/New York: Routledge. 38-71; Faarlund, Jan. T. 2004. The Syntax of Old Norse. Oxford: OUP. |
| Old Prussian | prg | | 0 | 1 | 0 | 1 | 0 | - | - | - | Schmalstieg, William. 1974. An Old Prussian Grammar. Pennsylvania State University: University Press; Siewierska, Anna. 1998. Constituent Order in the Languages of Europe. Berlin/New York: Mouton de Gruyter. |
| Oriya | ori | oya | 0 | 0 | 0 | - | 01 | 0 | - | 0 | Ray, Tapas S. 2003. Oriya. In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 487-522. |

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ossetic | oss | oss | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | |
| Panjabi ST | pan | pan | 0 | 0 | 0 | 0 | 0 | 0 | 01 | 0 | |
| Pennsylvania Dutch | pde | | 0 | 1 | 0 | 1 | 0 | 01 | 0 | 0 | Van Ness, Silke. 1994. Pensylvania German. In: König, Ekkehard and Johan van der Auwera. The Germanic Languages. London/New York: Routledge. 420-438. |
| Persian List | pes | prs | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | |
| Polish | pol | pol | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 01 | |
| Portuguese ST | por | por | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | Parkinson, Stephen. Portuguese. In: Harris, Martin and Nigel Vincent. The Romance Languages. London/Sydney: Coom Helm. 131-169. |
| Provencal | prv | | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | Wheeler, Max W. 1988. Occitan. In: Harris, Martin and Nigel Vincent. The Romance Languages. London/Sydney: Coom Helm. 246-278. |
| Riksmal | nor | nor | 0 | 1 | 0 | 01 | 0 | 1 | 0 | 0 | |
| Rumanian List | ron | rom | 1 | 1 | - | 1 | 0 | 1 | 0 | 0 | |
| Russian | rus | rus | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | |
| Sanskrit | san | | 0 | 01 | 0 | 0 | 0 | 0 | 01 | 0 | Ramat, Paolo and Anna Ramat. 1998. The Indo-European Languages. London/New York: Routledge; Bucknell, Roderick. 1996. Sanskrit Manual. A Quick-Reference Guide to the Phonology and Grammar of Classical Sanskrit. Dehli: Motilal Banarsidass; Whitney, William D. 1924. A Sanskrit Grammar. Including Both the Classical Language and the Older Dialects of Veda and Brahmana. Leipzig: Breitkopf & Haertel; Cardona, George. 2003. Sanskrit In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 115-178. |
| Sardinian C | sro | srd | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | Jones, Michael Allan. 1988. Sardinian. In: Harris, Martin and Nigel Vincent. The Romance Languages. London/Sydney: Coom Helm. 314-350. |
| Scots Gaelic | gla | gae | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| Serbocroatian | bos | scr | 0 | 1 | 0 | 01 | 0 | 1 | 0 | 0 | Browne, Wayles. Serbo-Croat. In: Comrie, Bernard and Greville G. Corbett. The Slavonic Languages. London/New York: Routledge. 306-387. |
| Sindhi | snd | | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | Kubchandani, Lachman M. 2003. Sindhi. In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 622-658; Yegorova, R.P. 1971. The Sindhi Language. Moscow: Nauka Gidwani; http://www.lmp.ucla.edu/Profile.aspx?LangID=201&menu=004 |
| Singhalese | sin | snh | 0 | 0 | 0 | 0 | 01 | 0 | 0 | 0 | Gair, James W. 2003. Sinhala. In: Cardona, George and Dhanesh Jain. The Indo-Aryan Languages. London/New York: Routledge. 766-817. |
| Slovak | slk | | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | Short, David. 1993. Slovak. In: Comrie, Bernard and Greville G. Corbett. The Slavonic Languages. London/New York: Routledge. 533-592. |
| Slovenian | slv | slo | 0 | 1 | 0 | 01 | 0 | 1 | - | 0 | Priestly, T.M.S. 1993. Slovene. In: Comrie, Bernard and Greville G. Corbett. The Slavonic Languages. London/New York: Routledge. 388-454. |
| Spanish | spa | spa | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 01 | |
| Swedish List | swe | swe | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | |
| Tadzik | tgk | taj | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | |
| Tocharian A | xto | | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | Winter, Werner. 1998. Tocharian. In: Ramat, Anna and Paolo Ramat. The Indo-European Languages. London/New York: Routledge.154-168; Xianlin, J., W. Winter and Pinault G-J. 1998. Fragments of the Tocharian A Maitreyasamiti-Nāṭaka of the Xinjiang Museum, China. Berlin/New York: Mouton de Gruyter. |
| Tocharian B | txb | | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | Gvozdanovic, Jadranka. 1992. Indo-European Numerals. Berlin/New York: Mouton de Gruyter; www.utexas.edu/cola/centers/lrc/eieol/tokol-0-X.html |
| Ukrainian | ukr | ukr | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | Pugh, Stefan and Ian Press. 1999. Ukrainian: A Comprehensive Grammar. London/New York: Routledge. |

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV | SOURCE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Urdu | urd | urd | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 | |
| Wakhi | wbl | wak | 0 | 01 | 0 | 0 | 0 | 0 | 0 | 0 | Gruenberg, Alexander L. 1988. La langue Wakhi. Paris. Maison des sciences de l'homme; Bashir, Elena. 2009. In: Windfuhr, Gernot. The Iranian Languages. London/New York: Routledge. 825-858. |
| Walloon | wln | | 01 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | Beardsmore, Hugo Baetens. 1971. Le Français Regional de Bruxelles. Presses Universitaires de Bruxelles; Valkhoff, Marius. 1938. Philologie et Littérature Wallonnes. Groningen: Wolters; Gaziaux, Jean-Jacques. 1987. Parler Wallon et Vie Rurale au Pays de Jodoigne. Publications Linguistiques de Louvain: Louvain-la-Neuve. |
| Waziri | pst | psh | 0 | 01 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Welsh N | cym | wel | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |

## S2.4 Uto-Aztecan data

| NAME | ISO | WALS | ADJ | ADP | DEM | GEN | NUM | OBV | REL | SBV |
|---|---|---|---|---|---|---|---|---|---|---|
| Aztec (Tetelcingo) | nhg | nht | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 01 |
| Aztec (Zacapoaxtla) | ncj | nhn | 01 | - | 0 | - | 0 | 1 | 01 | 0 |
| Cahuilla | chl | cah | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Chemehuevi | ute | cmh | 01 | 0 | - | 0 | 0 | 01 | 0 | 0 |
| Comanche | com | cmn | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cora | crn | cor | 1 | 0 | 0 | 01 | 0 | 1 | 0 | 1 |
| Eudeve | opt | eud | - | 0 | 0 | 0 | - | 0 | 0 | 0 |
| Guarijío | var | grj | 01 | 0 | 0 | 0 | 1 | 01 | - | 01 |
| Hopi | hop | hop | 0 | 0 | 0 | 0 | 0 | 0 | - | 0 |
| Huichol | hch | hui | - | - | - | 0 | - | 0 | 0 | 0 |
| Kawaiisu | xaw | kws | - | - | - | - | - | 01 | - | 01 |
| Luiseño | lui | lui | 1 | 0 | 0 | 0 | 0 | 01 | 0 | 0 |
| Mono | mnr | mno | - | 0 | 1 | 0 | - | 0 | 0 | 0 |
| Northern Paiute | pao | pno | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Northern Tepehuan | ntp | tpn | 01 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Opata | opt | eud | - | 0 | 0 | 0 | - | 0 | 0 | 0 |
| Pannamint | par | tsh | 0 | - | 01 | 0 | 0 | 0 | 01 | 0 |
| Papago-Pima | ood | ood | 0 | 01 | 0 | 01 | 0 | 1 | 0 | 1 |
| Pima de Onavas | pia | pba | 0 | - | 0 | 0 | 0 | 0 | 0 | 0 |
| Pipil | ppl | pip | 0 | 1 | - | 1 | 0 | 1 | 0 | 1 |
| Shoshoni (Gosiute Dialect) | shh | sho | 0 | - | 0 | - | - | 0 | - | 0 |
| Southeastern Tepehuan | stp | tps | 1 | 0 | 0 | - | - | 1 | - | 1 |
| Southern Paiute | ute | cmh | 01 | 0 | - | 0 | 0 | 01 | 0 | 0 |
| Southern Ute | ute | ute | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 01 |
| Tarahumara | tar | tce | 0 | 0 | 0 | - | - | 0 | 0 | 0 |
| Yaqui | yaq | yaq | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

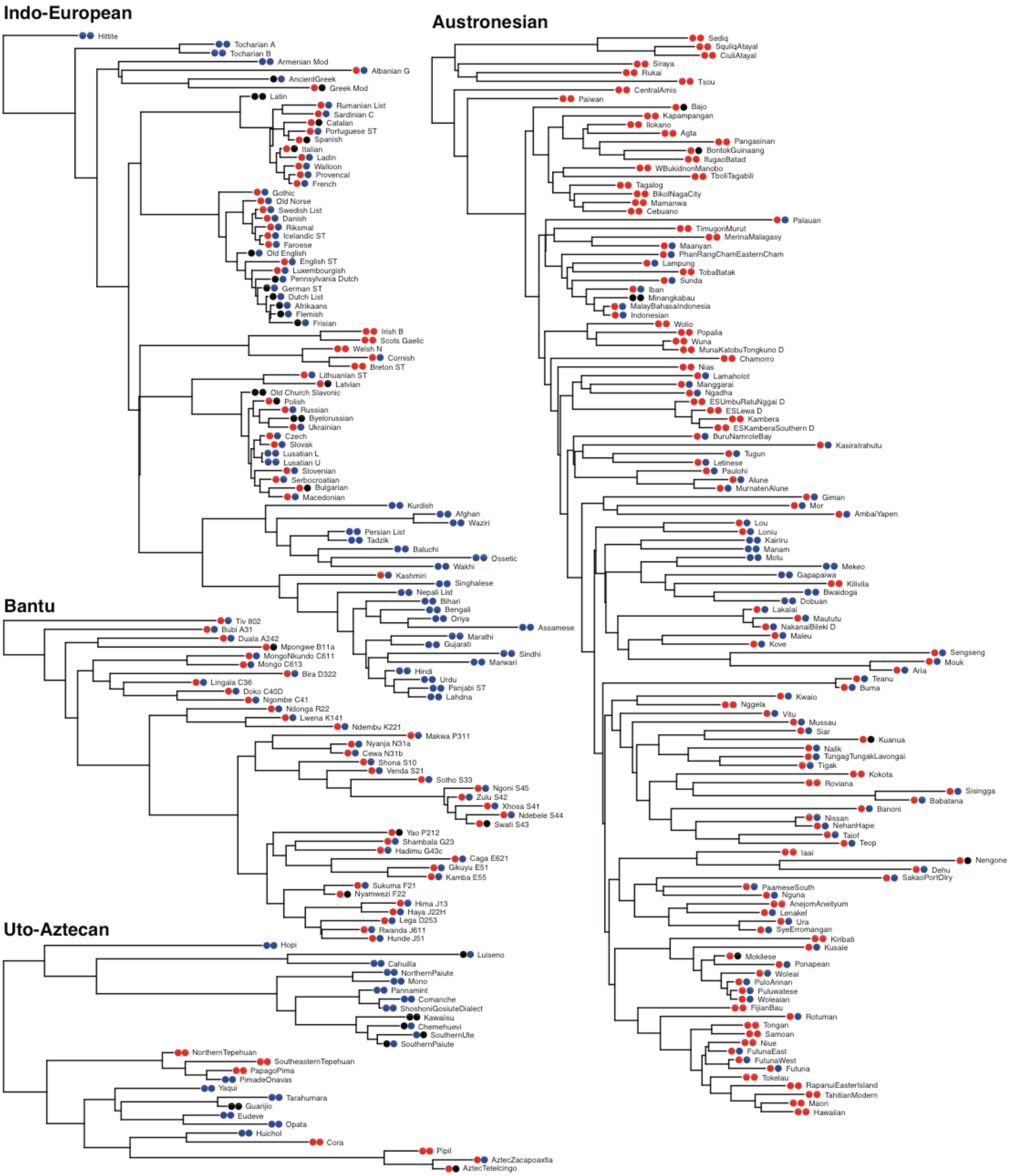


*Figure S3: Character states for Subject~Verb order and Object-Verb order. Blue-blue indicates SV/OV; red-red indicates VS/VO; blue-red indicates SV/VO; red-blue indicates VS/OV. Polymorphic characters are in black. There is strong evidence for dependency (BF=13.7) in the Uto-Aztecan lineage only.*

## *S3 Numerical results*

This table shows the median Bayes Factors for each of the possible dependencies. Positive values from 2-5 indicate weak support for the 'dependent' hypothesis over the independent hypothesis. Values from 5 are conventionally considered 'strong' evidence (marked boldface in the results for each language family). Predicted dependencies according to Dryer are indicated by boldface labels of the feature pair; Greenbergian universals pertaining to particular feature pairs are indicated by their number after the label. Correlations involving invariant states (in Bantu) were not tested. The Dryer and Greenberg predictions are compared in Figure S4. AN=Austronesian, BA=Bantu, IE=Indo-European, UA=Uto-Aztecan. ADJ=Order of Adjective and Noun; ADP=Adposition Order; DEM=Demonstrative and Noun; GEN=Genitive and Noun; NUM=Numeral and Noun; OBV=Order of Object and Verb; REL=Relative clause and Noun; SBV=Subject and Verb.

| | AN | BA | IE | UA |
|---|---|---|---|---|
| ADJ-ADP | 0.99 | - | 3.95 | 2.78 |
| ADJ-DEM (18) | **5.53** | 2.9 | **7.64** | 2.51 |
| ADJ-GEN (5) | 2.83 | - | **21.23** | 3.04 |
| ADJ-NUM (18) | **15.64** | 3.99 | 0.56 | 1.49 |
| ADJ-OBV (5,17) | 1.21 | - | 1.43 | 2.31 |
| ADJ-REL (24) | **5.33** | -0.58 | -0.25 | **5.02** |
| ADJ-SBV | -1.09 | -0.64 | **6.88** | 1.97 |
| ADP-DEM | -1.81 | 4.11 | 3.16 | -1.13 |
| **ADP-GEN** (2) | 3.74 | - | **13.65** | 3.21 |
| ADP-NUM | **7.26** | - | -0.6 | -1.41 |
| **ADP-OBV** (3,4) | **15.34** | - | **13.34** | 3.82 |
| **ADP-REL** (24) | -1.79 | - | 2.51 | -0.45 |
| ADP-SBV | 2.87 | - | 3.39 | 2.56 |
| DEM-GEN | -2.26 | - | **7.03** | -0.54 |
| DEM-NUM | -2.77 | **5.01** | 0.79 | 0.09 |
| DEM-OBV | -2.53 | **-** | **7.55** | -0.08 |
| DEM-REL | 3.28 | - | -3.86 | 1.76 |
| DEM-SBV | -3.69 | -0.12 | 3 | 2.26 |
| GEN-NUM | **18.26** | 0.33 | 1.66 | -2.68 |
| **GEN-OBV** | -0.88 | - | **5.27** | **5.22** |
| **GEN-REL** | 3.04 | - | 1.05 | 0.09 |
| GEN-SBV | 1.15 | - | **5.14** | 3.22 |
| NUM-OBV | 1.68 | - | -1.36 | 0.52 |
| NUM-REL | 1.09 | -0.89 | 1.24 | -1.05 |
| NUM-SBV | -0.92 | -0.72 | 0.42 | 1.1 |
| **OBV-REL** (13) | -1.01 | - | 1.56 | -1.52 |
| OBV-SBV (1) | 3.93 | - | 4.93 | **13.57** |
| REL-SBV | 3.23 | - | -4.7 | 0.34 |

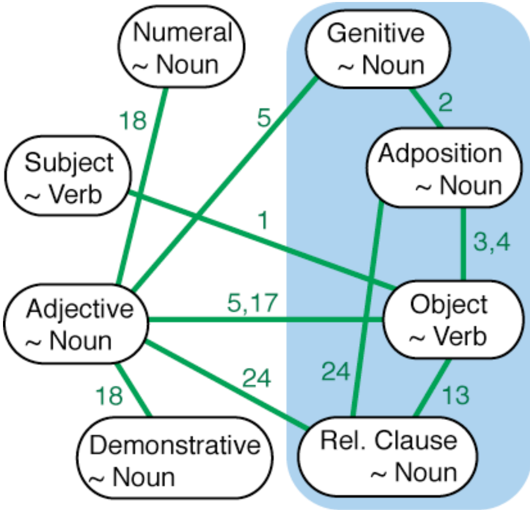


*Figure S4: This figure compares Dryer's claims about word order dependencies (features in the blue area) to dependencies predicted from Greenberg's universals (green; the number of the universal according to Greenberg's classification is marked on the line).*

## S4 References


Bastin, Y., Coupez, A. & Mann, M. (1999) Continuity and divergence in the Bantu languages: perspectives from a lexicostatistic study. *Annales, Sciences Humaines*, 162.

Davis, G. (2002) Syntax of West Germanic: The syntax of Old English and Old High German. Kummerle.

Dryer, M.S. (2005a) Determining the Order of Subject, Object, and Verb. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Dryer, M.S. (2005b) Order of Object and Verb. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Dryer, M.S. (2005c) Determining Dominant Word Order. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Dryer, M.S. (2008a) Order of Subject and Verb. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 82. Available online at http://wals.info/feature/82. Accessed on 2009-03-29.

Dryer, M.S. (2008b) Order of Object and Verb. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 83. Available online at http://wals.info/feature/83. Accessed on 2009-03-29.

Dryer, M.S. (2008c) Order of Adposition and Noun Phrase. In: Haspelmath, M., Dryer, M.S., Gil, D., & Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 85. Available online at http://wals.info/feature/85. Accessed on 2009-03-29.

Dryer, M.S. (2008d) Order of Genitive and Noun. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 86. Available online at http://wals.info/feature/86. Accessed on 2009-03-29.

Dryer, M.S. (2008e) Order of Adjective and Noun. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 87. Available online at http://wals.info/feature/87. Accessed on 2009-03-29.

Dryer, M.S. (2008f) Order of Demonstrative and Noun. In: Haspelmath, M., Dryer, M.S., Gil, D. &

Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 88. Available online at http://wals.info/feature/88. Accessed on 2009-03-29.

Dryer, M.S. (2008g) Order of Numeral and Noun. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 89. Available online at http://wals.info/feature/89. Accessed on 2009-03-29.

Dryer, M.S. (2008h) Order of Relative Clause and Noun. In: Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 90. Available online at http://wals.info/feature/90. Accessed on 2009-03-29.

Durie, M. & Ross, M. (1996) *The Comparative Method Reviewed: Regularity and irregularity in language change*. Oxford: Oxford University Press.

Dyen, I., Kruskal, J.B. & Black, P. (1992) An Indo-European classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society,* 82:iii-132.

Edwards, A.W.E. (1972) *Likelihood*. Cambridge: Cambridge University Press.

Embleton, S.M. (1986) *Statistics in Historical Linguistics*. Bochum: Studienverlag Brockmeyer.

Gray, R.D. & Atkinson, Q.D. (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature,* 426:435-439.

Gray, R.D., Drummond, A.J., & Greenhill, S.J. (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science,* 323:479-483.

Greenhill, S.J., Blust, R, & Gray, R.D. (2008) The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics. *Evolutionary Bioinformatics*, 4:271-283.

Greenhill, S.J., Currie, T.E., & Gray, R.D. (2009) Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society B*, *276*, 2299-2306.

Haspelmath, M., Dryer, M., Gil, D. & Comrie, B. (2005) *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Hunley, K., Dunn, M., Lindström, E., Reesink, G., Terrill, A., Healy, M.E., Koki, G., Friedlaender, F.R., Friedlaender, J.S. (2008) Genetic and Linguistic coevolution in Northern Island Melanesia. *PLoS Genet*, 4: e1000239.

Kayne, R.S. (1994) *The Antisymmetry of Syntax.* Linguistic Inquiry Monograph 25. MIT Press.

Lansing, S.J., Cox, M.P., Downey, S.S., Gabler, B.M., Hallmark, B., Karafet, T.M., Norquest, P., Schoenfelder, J.W., Sudoyo, H., Watkins, J.C., Hammer, M.F. (2007) Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences,* 104:16022-16026.

Mace, R. & Pagel, M. (1994) The comparative method in anthropology. *Current Anthropology,* 35:549-564.

Mayo, P. (1993) Belorussian. In: Comrie, B. & Corbett, G.G. (eds.) *The Slavonic Languages*.

Pagel, M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B,* 255:37-45.

Pagel, M. & Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology,* 53:571-81.

Pagel, M. & Meade, A. (2005) Mixture models in phylogenetic inference. In O. Gascuel (Ed.), *Mathematics of Evolution and Phylogeny* (pp. 121-142). Oxford: Oxford University Press.

Pagel, M. & Meade, A. (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist*, 167: 808-825.

Penny, D., McComish, B.J., Charleston, M.A. & Hendy, M.D. (2001) Mathematical elegance with biochemical realism: The covarion model of molecular evolution. *Journal of Molecular Evolution*, 53:711-723.