

Grammar and Phylogenies

Robin J. Ryder

September 15, 2006

Abstract

Using exclusively syntactical features, we attempt to reconstruct language evolution trees using methods from molecular biology: a maximum parsimony reconstruction, and a maximum likelihood estimator using Bayesian Markov chain Monte Carlo. We first check how well these methods work on languages for which classification by the traditional comparative method is uncontroversial, both at the genus and family level. After refining the data we use, so as to incorporate only slowly-evolving features, and building a simple model of evolution of syntactical features, we find that the maximum likelihood method is efficient at determining phylogenies. We then apply this to some controversial classifications of languages of Asia, and find little or no evidence for them.

And the whole earth was of one language, and of one speech.
(Genesis 11:1)

Contents

1	Introduction	3
1.1	Languages evolve	3
1.2	Similarities between language and genes	3
1.3	Application of phylogenetics to language trees	4
2	Grammar evolution	4
3	Results	7
3.1	Indo-European languages	7
3.1.1	Maximum parsimony trees	8
3.1.2	Bayesian analysis	11
3.2	Languages from different families	12
3.2.1	Maximum parsimony trees	12
3.2.2	Bayesian analysis	13
3.2.3	Refining the features	13
4	Improving the distance	15
4.1	Results	16
4.1.1	Languages from different families	16
4.1.2	Languages of Eurasia	16
5	Conclusions	17
6	Acknowledgements	18
A	Description of the languages	19
B	Details of the features	21

1 Introduction

1.1 Languages evolve

It is generally accepted that most languages in Europe and many languages in India stem from a common ancestor, between 6000 and 8000 years before present (BP). These languages are called Indo-European languages, and their ancestor, Proto-Indo-European. The evidence for this is mainly in similarities between words. For example¹:

English	Latin	Greek	Sanskrit	Avestan	Tocharian A
father	pater	paté:r	pitár-	pitar-	pa:car
mother	ma:ter	má:te:r	ma:tar-	ma:tar-	ma:car
foot	pedus	podós	pá:t, pá:dam	pad-	päts
salt	sa:l	háls	sal-ilá		sa:le
young	iuvenis		yúva	yavan	
red	rūbidos	ereuthos	róhita-	raoiđita-	rätr-ārkyant

(Data from Justus (2006))

Not only are the similarities numerous, the differences also exhibit very regular patterns; for example, a *p* in Sanskrit will become an *f* in English (*father/pitár-*, *foot/pá:t*). These correspondences are too numerous to be a coincidence, so these languages must be related. (Joseph, 2001; Rosenfelder, 2002)

The field of comparative linguistics deals with finding these similarities, and inferring the relationships between languages. For the recent history, written records help document very accurately the changes, so it is clear that French, Italian, Spanish and Romanian are all closely related, and that Latin is their common ancestor. For deeper relationships, such as between different branches of Indo-European languages, the topology of the tree is less clear, and dating the nodes is even harder.

Comparative linguists often view their subject more like an art than a science. Until very recently, no systematic method had ever been applied to language evolution.

1.2 Similarities between language and genes

Darwin (1871) already noted that language evolution is similar to biological evolution. Though there is no notion of fitness in languages, the process of small changes accumulating until two different species, or languages, emerge, is the same. Languages borrow lexical items, sounds and syntactic features from one another, in a process similar to the infection of a genome by a bacteria. For a long time, it has seemed natural to represent relationships between languages on a tree, as for biological species.

Representing language evolution on a tree only captures part of the phenomenon. Borrowing of words or grammatical features is not represented on a tree, and can lead to erroneous classifications if borrowing is not recognized as such (McMahon and McMahon, 2005). Nonetheless, we will consider only tree models, which are much simpler, and make clear a major component of the history of languages.

¹Appendix A contains a short description of all the languages mentioned.

The terminology in language classification is similar to that for species: languages are grouped into *genera*, and several genera form a *family*. Other levels of classification include *subfamilies*, *subgenera* and *superfamilies*. It is believed that all languages in one branch have a common ancestor, at the root of the tree. For example, all Germanic languages (English, German, Dutch, Swedish, Danish...) stem from Proto-Germanic, which was spoken about 1800 years BP; all Indo-European languages stem from Proto-Indo-European. Some linguists believe that all human languages share a common ancestor, called the Proto-World language. While groupings into families can be uncontroversial, there is no consensus on the relationship between language families. There is so far little evidence for a Proto-World language, and nothing substantial can be inferred about it. (Rosenfelder, 2006)

1.3 Application of phylogenetics to language trees

Recently, efforts have been made to adapt phylogenetic algorithms to language data. Given the similarity between language evolution and biological evolution, and the common tree-like representations, it seems plausible that the methods which have been proved to be efficient in biology could yield some interesting results in language evolution, and provide a quantitative description, rather than the qualitative description provided until now by historical linguists. The amount of data recently acquired for hundreds of languages also makes this approach much more feasible nowadays than it used to be.

Gray and Jordan (2000) were among the first to apply the phylogenetic method to linguistic data. Using 5,185 lexical items for 77 Austronesian languages, they were able to obtain a single most parsimonious tree. This was an argument in favour of the family-tree model for languages. They claimed it also supported one of two main theories on expansion in Austronesia around 6000 BP.

Gray and Atkinson (2003) also applied the phylogenetic method to Indo-European languages. The results obtained, using penalized maximum likelihood, were very close to the consensus tree given by classical comparative linguistics methods (Fig. 1). The genera yielded by their analysis correspond exactly to those usually accepted by historical linguists. Several of the subfamilies they found also corresponded to what was already believed, such as a subfamily grouping Germanic, Romance and Celtic languages. Other parts of the tree were unresolved, such as the position of Albanian. While their results do not show any new groupings, the fact that the results correspond to what is accepted by linguists gave hope that phylogenetics could be successfully used in other, less studied, language families.

McMahon and McMahon (2005) also applied the Neighbour-joining method to 95 Indo-European languages and 200 lexical items. The unrooted tree they obtained showed the ten genera generally admitted for Indo-European languages, but no clear relationship appeared between the genera.

2 Grammar evolution

Comparative linguistics uses mainly lexical items to build trees. However, syntactical features evolve in a similar fashion. Features of interest include, among

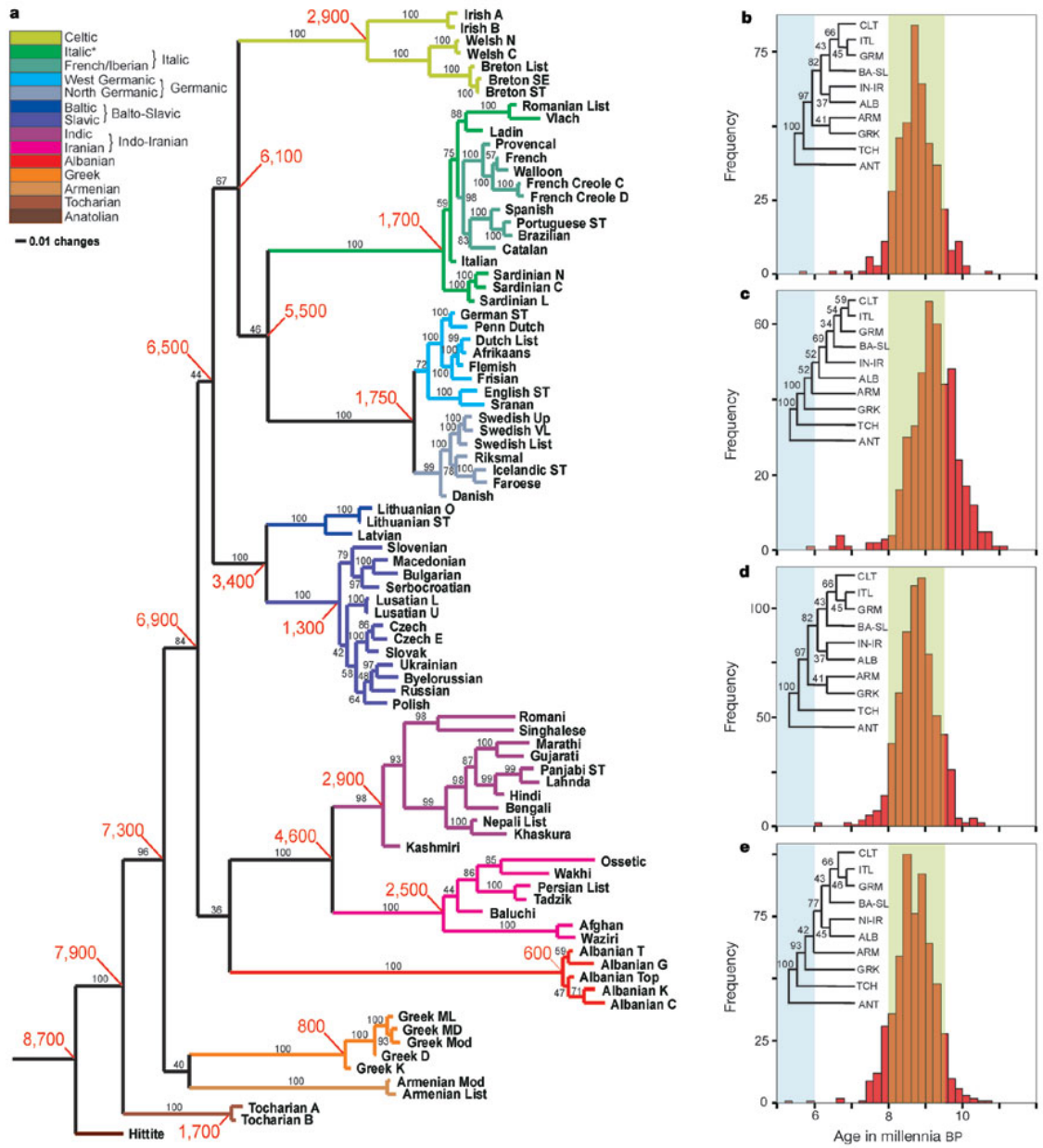


Figure 1: Consensus tree of Indo-European languages obtained by Gray and Atkinson (2003) using penalized maximum likelihood on lexical items.

others, the presence and characteristics of declensions, word order, coding of tenses and moods and gender systems.

There are four possible reasons why two languages share a syntactical trait (Comrie, 1989):

1. Chance: given that each trait can only take a finite number of states, unrelated languages will share features. For example, the Mbabaram language of Australia uses the word *dog* for a dog, but it has been shown that is purely a coincidence (Trask, 1996).
2. Genetic relationship
3. Borrowing
4. Language universal: some features are common to all human languages

Given the large number of features we will be looking at, repeated coincidences are highly unlikely. The main issue is to know how important genetic relationship and borrowing are, and whether we can use grammatical features to infer language history.

Nichols (1999) argues that grammatical features evolve rapidly, and are subject to regional influences; on the other hand, Dunn et al. (2005) argue that grammatical features evolve much more slowly than lexical items, and that they can be used to calculate phylogenies further than the time barrier imposed by semantics (about 8000 years). Indeed, while some features exhibit a great deal a variance across languages, features such as the use of prepositions or postpositions are extremely stable.

According to Comrie et al. (1987), word order is easier to borrow than basic vocabulary (words such as *mother*, *one* or *head*), but harder than cultural vocabulary (such as words related to technology). On the contrary, Thomason and Kaufman (1988) attempted to create a borrowing scale, which can be reinterpreted as:

- Lexical items
- Function words, minor phonological, syntactic, and lexical semantic features
- Adpositions, derivational affixes, phonemes
- Word order, distinctive features in phonology, inflectional morphology
- significant typological disruption, phonetic changes

with category 1 being the easiest to borrow, and category 5 the hardest (Matras, 2000).

In this report, we use the syntactical features described in the *World Atlas of Language Structures* (Dryer et al., 2005), hereafter *WALS*. *WALS* lists 106 syntactical features for 2,559 languages. These features are all discrete, with 2 to 9 possible values each. A complete list is available in appendix B. It also lists some phonological and lexical features, which we did not use in this study.

Several features show great stability within the Indo-European languages. The following 20 features are identical in all Indo-European families, except for up to two languages or genera. Other languages in the region do not share

these traits. This is a strong indication that these values come from the ancestor language, Proto-Indo-European.

- 20:1** Grammatical markers are always bound to a host word
- 24:2** In Possessive noun phrases, the possessor is dependent-marked (except in Persian and Greek)
- 28:3** Inflectional case marking is syncretic for core and non-core cases (except in Persian and English)
- 33:2** Nominal plurality is coded with a plural suffix (except in Maithili and Celtic languages)
- 34:6** Plural occurs in all nouns and is always obligatory (except in Indic languages)
- 39:3** No Inclusive/Exclusive opposition in independent pronouns
- 58:2** No obligatorily possessed nouns (except in Ossetic)
- 59:2** Two possessive classes (except in Ossetic)
- 63:1** *and* and *with* are not identical
- 66:1** Past/non-past distinction marked; no remoteness distinction (except in Baluchi)
- 74:1** It is possible to express situational possibility with affixes on verbs (except in Gojri)
- 80:1** No singular-plural pairs
- 89:1** Numeral precedes Noun (except in Indic languages)
- 99:2** Alignment of case-marking Pronouns is standard Nominative-Accusative (except in Hindi)
- 100:6** Alignment of verbal person marking is split (except in Kashmiri)
- 105:4** Ditransitive constructions are mixed (except in English)
- 107:1** There is a passive construction
- 108:3** There is no antipassive construction
- 109:8** There is no applicative construction
- 118:2** Predicative adjectives have nonverbal encoding

Even though there is no consensus on the speed of grammar evolution, this list seems to show that at least some syntactical features are retained from the ancestor; it is therefore probably worth to try and infer phylogenies from grammatical features. If the results are similar to those yielded by phylogenetic methods using lexical items and the traditional comparative method, this will be further evidence for those classifications. If the results differ, it would be interesting to study the reasons for the divergence.

3 Results

3.1 Indo-European languages

We first looked at languages for which a lot of evolutionary history is known: Indo-European languages. While it would have been better to sample languages at random, many languages have mostly missing data; Stoneking (2006) estimated that 84 % of the data was missing in WALS. The following languages were therefore chosen for having at least 40 data points, and for representing all the living Indo-European genera:

- Germanic languages: Dutch, English, German (West Germanic); Danish, Icelandic (North Germanic)
- Romance languages: French, Italian, Romanian, Spanish
- Celtic languages: Breton, Irish, Welsh
- Baltic languages: Latvian, Lithuanian
- Slavic languages: Belarusian, Bulgarian, Czech, Macedonian, Polish, Russian, Serbo-Croatian, Slovak, Slovenian, Ukrainian
- Indo-Iranian languages: Hindi, Marathi, Panjabi (Indo-Aryan); Kashmiri (Dardic)
- Albanian
- Armenian
- Greek

While the classification of these languages into genera is uncontroversial, the relationship between the genera is unclear. The classification of these languages, and the dating of the nodes, has been called "the most intensively studied, yet still most recalcitrant, problem of historical linguistics" (Diamond and Bellwood, 2003). It should also be noted that five of these languages form the so-called Balkan Sprachbund: Albanian, Bulgarian, Greek, Macedonian and Romanian. Though they belong to four different genera, these languages have been in very close contact for several centuries, and therefore share many lexical, syntactical and phonological features (Trask, 1996).

3.1.1 Maximum parsimony trees

We first searched for the maximum parsimony tree for the data; Ringe et al. (2006) estimated that maximum parsimony is the most effective phylogenetic method for lexical data, though they did not consider any sort of Bayesian analysis. The distance we used is naive: we gave the same weight to all features; within a feature, we estimated that all the values were equidistant. This is refined in later sections. We used the `pars` package of the PHYLIP software (Felsenstein, 2005), which calculates the most parsimonious tree(s) for discrete data. The most parsimonious tree is defined as the one requiring the least changes overall; in our case, the data is discrete, unordered, and multi-state (up to 9 states per feature).

The assumptions of this method are (Felsenstein, 2004):

1. Ancestral states are unknown.
2. Different characters evolve independently.
3. Different lineages evolve independently.
4. Changes to all other states are equally probable (Wagner).
5. These changes are *a priori* improbable over the evolutionary time spans involved in the differentiation of the group in question.

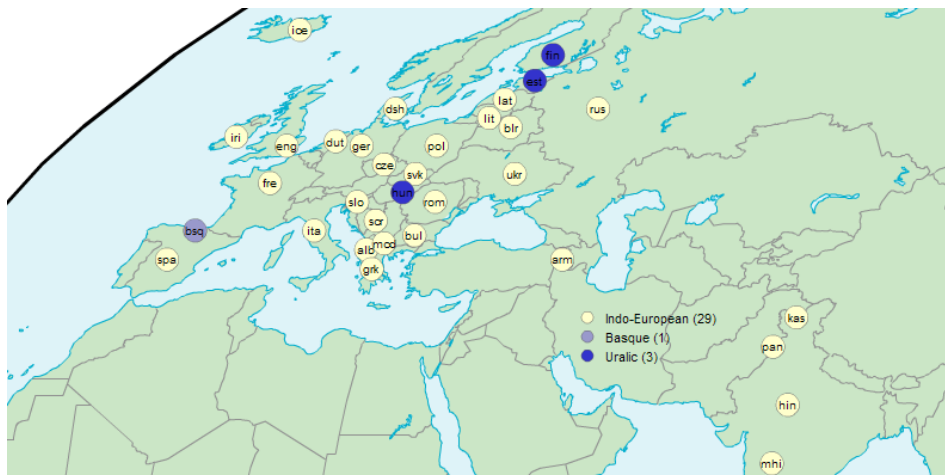


Figure 2: *Map of Indo-European languages used. Made using the electronic version of Dryer et al. (2005).*

6. Other kinds of evolutionary event such as retention of polymorphism are far less probable than these state changes.
7. Rates of evolution in different lineages are sufficiently low that two changes in a long segment of the tree are far less probable than one change in a short segment.

These assumptions are not all justified: lineages do not evolve independently (assumption 3), since borrowing occurs between different lineages. Similarly, all characters are not independent (assumption 2): for example, a language with a Subject-Verb-Object basic word order is more likely to use prepositions (adposition before the noun, as in *over the world*) than postpositions (adposition after the noun, as in *the world over*) (Dryer, 2005). As stated above, the assumption that within a feature, all values are equidistant, is naive (assumption 4); this is dealt with in section 3.1.2.

While borrowing can sometimes be easily detected for lexical items, it is much less obvious for syntactical features. It would be possible to select a much reduced sample of independent features, but there would then not be enough data to construct a tree. Furthermore, it is also when languages are exceptions to a correlation that they can easily be detected as being cousins: Finnish and Estonian are in the same genus, and are both in the 3% of languages with Subject-Verb-Object word order and postpositions. We therefore used the method even though not all assumptions are justified.

To test whether geographical contact was likely to disrupt the tree, four non-Indo-European languages of European were added: Basque, which is a language isolate ², and Finnish, Estonian and Hungarian, which are in the Finno-Ugric family. A map of the languages is given in Fig. 2.

If languages retain a significant amount of the syntactic features of their ancestors, we expect the languages to be clustered according to their genus, and

²A language isolate is one with no (known) cousins.

the four non-Indo-European languages to be outliers to the tree. On the other hand, if geographical contact is more important, we should observe similarities between languages which are not related genetically, but are in close contact: Basque, Spanish and French; Hungarian and Slavic languages; Estonian and Baltic languages.

Over 100 most parsimonious trees were found, but all were very similar to Fig. 3.

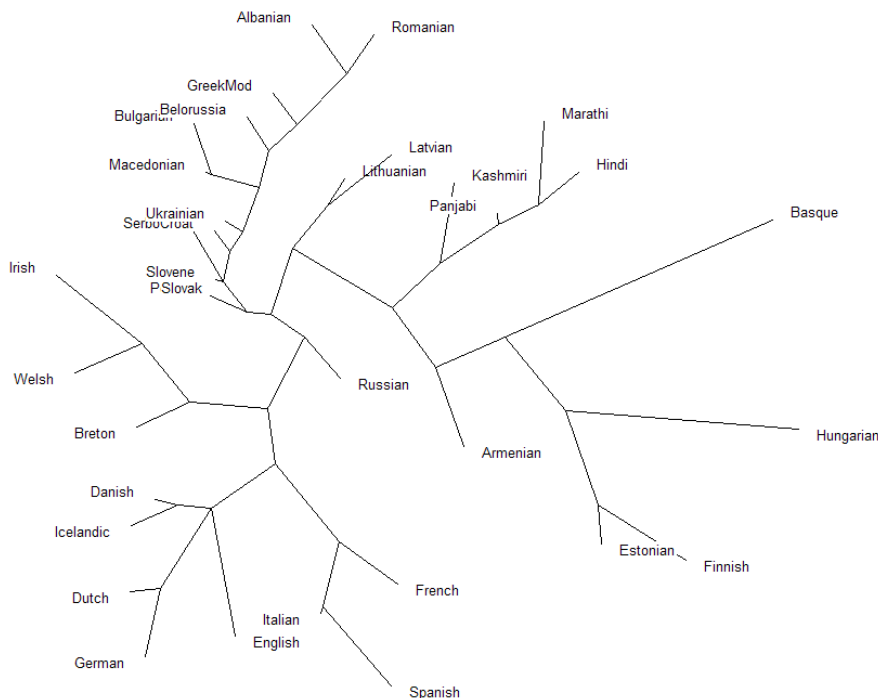


Figure 3: *Typical most parsimonious tree for Indo-European languages, using the pars package of PHYLIP. Edited with TreeView (Page, 1996).*

The most parsimonious tree shows that genetic relationship is much stronger than geographical contact: almost all the languages are clustered into the correct genus. The two exceptions are Russian, a slight outlier to the Slavic genus, and the Balkan Sprachbund languages, which are all grouped together within the Slavic group. The fact that they are grouped together shows that geographical contact can induce large amounts of borrowing of syntactical features; the fact that they are clustered with the Slavic languages is probably a consequence of there being two Slavic languages in the Balkan Sprachbund (Bulgarian and Macedonian), while the other three are alone in their respective genera.

The Germanic, Romance (with the exception of Romanian), Celtic, Baltic and Indo-Iranian genera are reconstructed successfully. The consensus between historical linguists is that the Romance, Germanic, and Celtic languages are closely related (Gray and Atkinson, 2003); this tree captures that as well.

Interestingly, this method seems to capture the deep relationships better than the shallow: while the genera are correctly reconstructed, the topology within the genera is at times random. For instance, Panjabi is a closer cousin

to Hindi and Marathi than Kashmiri is; Breton is a closer cousin to Welsh than Irish is. It therefore seems that syntactical features can be used to group languages into genera, and possibly into families or subfamilies, but not to study the recent history of the language.

3.1.2 Bayesian analysis

Work by Wichmann and Saunders (2006ms) showed that a Bayesian analysis can be much more efficient than the maximum parsimony method. We tried to infer the topology of the tree using Bayesian analysis. We used Metropolis-coupled Markov Chain Monte Carlo, or (MC)³, to explore the posterior distribution in a maximum-likelihood estimator model. We used a uniform prior on the topology of the tree. Several priors for the evolution rates were tried; the results showed no dependency on the prior. The results shown are those for an inverse-gamma prior. For now, the transition matrix we used is very simple: for each feature, all the transition rates are equal. We do not assume any ordering between different values of a feature.

(MC)³ runs n chains, of which $n - 1$ are heated. A heated chain has steady-state distribution $\pi_i(X) = \pi(X)^{\beta_i}$ with $\beta_i = \frac{1}{1+T(i-1)}$, where π is the distribution we want to evaluate, i the number of the chain, and T is called the temperature. After each iteration, an attempt is made to swap two randomly chosen chains i and j ; the swap is accepted with probability

$$\min \left(1, \frac{\pi_i(X_t^{(j)})\pi_j(X_t^{(i)})}{\pi_i(X_t^{(i)})\pi_j(X_t^{(j)})} \right).$$

Inferences are based only on the "cold" chain ($i = 1, \beta_i = 1$). The heating flattens the steady-state distribution, and improves the mixing by making it easier to go from one mode to another (Gilks and Roberts, 1996). We used $n = 4$ and $T = 0.20$.

We ran two simultaneous independent analyses. Convergence was assessed using the average standard deviation of split frequencies between the two analyses; convergence was assumed when the average standard deviation of split frequencies was under 0.01. The first 25% of the run were discarded as a burn-in.

For computational reasons, we used only 26 languages. In the results, almost all the true genera had posterior probability 1.00:

Branch	Posterior Probability
Indo-Iranian	1.00
Slavic	1.00
Germanic	1.00
Celtic	1.00
Italo-Western Romance	1.00
Armenian and Indo-Iranian	.998
Albanian, Romanian and Greek	.97

Only one genus was not correctly reconstructed: Romanian was not recognized as a Romance language; it was instead attracted by the Balkan Sprachbund (the Baltic genus is not included because our smaller subset only included one Baltic language). Overall, this method outperformed the maximum parsimony method: all the Slavic languages were grouped together, and the Balkan

Sprachbund was separated from the slavic genus. Also, the maximum parsimony method was not good for identifying relationships within genera; this method identified many subgroupings, *e.g.* Indo-Aryan within Indo-Iranian (posterior probability: 0.48). The maximum parsimony tree supported a grouping of Germanic and Romance, then Celtic; our Bayesian analysis supports a grouping of Romance and Celtic first (posterior probability: 0.67). Both hypotheses have been supported by historical linguists using the traditional comparative method.

Interestingly, Armenian was grouped in a subfamily with the Indo-Iranian languages with posterior probability very close to 1; this could be due to borrowing induced by extended contact.

3.2 Languages from different families

In many cases, such as native North American languages, classification into families is unclear. If possible, it would therefore be interesting to be able to use grammatical features to classify these languages. The rationale for doing this is that while some grammatical features evolve very rapidly, others seem to evolve very slowly, and should therefore retain a lot of information about their ancestor.

29 languages from 10 different families were selected, once again partly for the large amount of data available, and partly for their spread across different families and continents:

Family	Languages
Arawakan	Apuriña
Sino-Tibetan	Bawm, Burmese, Garo, Meithei
Macro-Ge	Canela-Kraho
Australian	Gooniyandi, Maranungku, Martuthunira, Nunggubuyu, Waradaman
Eskimo-Aleut	Greelandic, Yup'ik
Afro-Asiatic	Hausa, Kera, Oromo, Somali
Nakh-Daghestan	Hunzib, Ingush, Lezgian
Niger-Congo	Igbo, Luvale, Nkore-Kiga, Sango Swahili
Khoisan	Ju 'hoan, Khoelkhoe
Hokan	Maricopa, Pomo

3.2.1 Maximum parsimony trees

We first used PHYLIP to try and cluster these languages into families using maximum parsimony. The algorithm attempts to put all the languages on the same tree, giving the illusion that it might be finding relationships between families, and discovering superfamilies. In fact, the likelihood of these groupings is very small, and should not be interpreted as a result.

Fig. 5 shows one of the two most parsimonious trees. Some families are correctly grouped together: Eskimo-Aleut, Nakh-Daghestan, Niger-Congo, and most of Sino-Tibetan and Australian. On the other hand, the algorithm does not find the Afro-Asiatic, Khoisan and Hokan families. Overall, the results are rather poor.

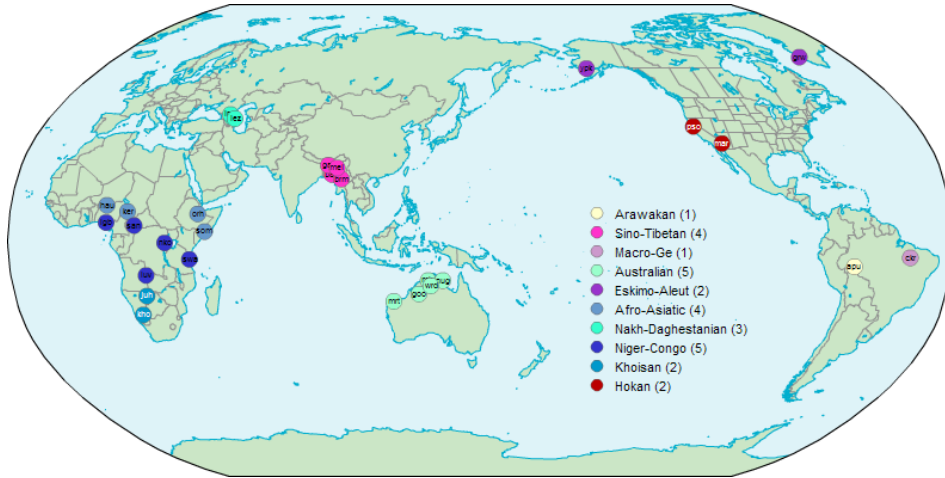


Figure 4: Map of the set of 29 languages from different families, shown by family. Made using the electronic version of Dryer et al. (2005).

3.2.2 Bayesian analysis

Using Bayesian analysis, we are able to estimate the probability of different groupings, even if these groupings do not appear on the most parsimonious tree.

The following groupings were found with high posterior probability (over 90%):

Branch	Posterior probability
Nakh-Daghestan	1.0
Bantoid	.99
Niger-Congo	.95
3 Australian languages	.95

Several false groupings were predicted with a posterior probability of up to 66% (Meithei, a Sino-Tibetan language, was grouped with Pomo, a native American language, with posterior probability 66%). The other families were found with lower posterior probabilities (Eskimo-Aleut: 51%), or sometimes were not found at all (Khoisan, Hokan, though these families are probably the most controversial amongst those we selected) (by not found, we mean that the posterior probability was under 5%).

The algorithm also found (with posterior probability 69%) a large superfamily containing languages of Eastern Asia, North America and Australia. While it is possible that all these languages have a common ancestor, the fact that several much closely related families were not found means that this is probably little more than a coincidence.

3.2.3 Refining the features

Given the limited success of our approach in the previous section, we tried to refine the features: it is clear that some features will evolve much more

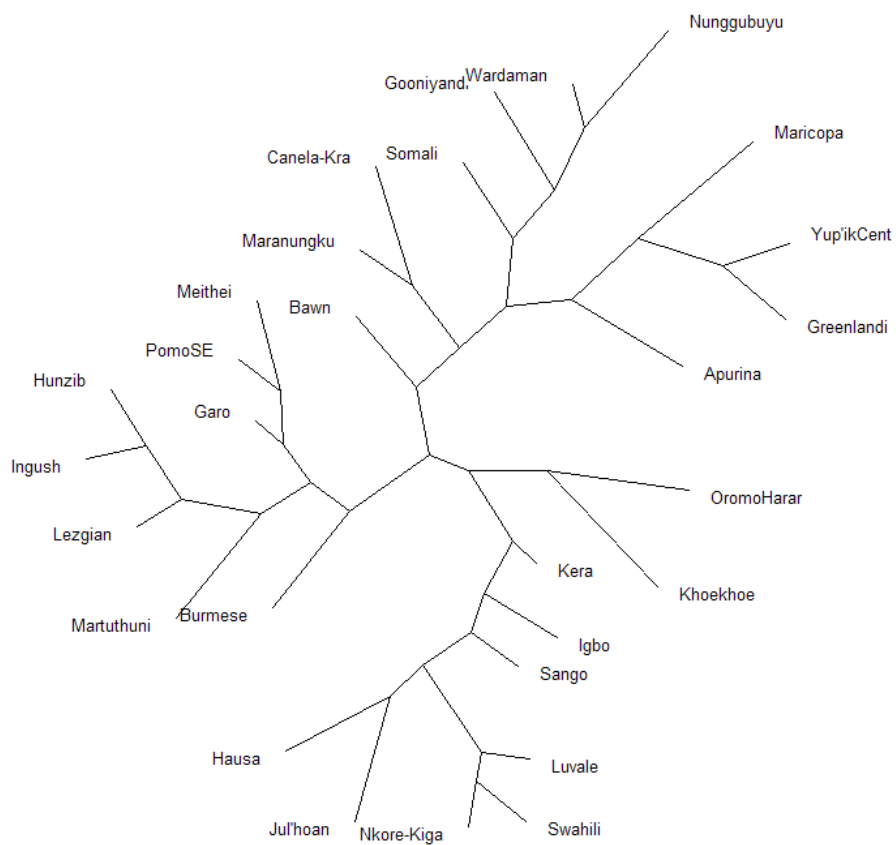


Figure 5: *One of two most parsimonious trees for languages from different families, as given by the `pars` package of PHYLIP. Edited with TreeView (Page, 1996).*

rapidly than others, so we attempted to identify those who evolve slowly, and are therefore likely to reflect the ancestor language.

Using a different set of 27 languages from 8 families, we looked for features which were identical in all the languages of at least half of the families. This gave us a list of 52 slowly-evolving features. We then used only these features in a maximum likelihood analysis; the following branches were found:

Branch	Posterior probability
Nakh-Daghestanian	.999
Gunwinygguan	.99
Eskimo-Aleut	.96
Bantoid	.76
Igbo and Sango	.73
Gooniyandi and Martuthunira	.64
Hokan	.61
Bawn and Burmese	.55
Niger-Congo	.51

More families or subfamilies were found this time; no bogus grouping had a posterior probability above 50% (the highest one was Ju|'hoan with Nkore-Kiga, 45%). The Khoisan family was still not found. Several groupings which could be subfamilies were also found: Igbo and Sango (two Niger-Congo languages), Gooniyandi and Martuthunira (two Australian languages), Bawn and Burmese (two Tibeto-Burman languages). The superfamily of Asian, Australian and North American languages disappeared, which confirms that it was found as a result of a coincidence, or because the model was too poor.

4 Improving the distance

Until now, the distance between two values for a features was always 1. This makes little sense: for example, feature 79 deals with "Suppletion according to tense and aspect". English is said to have suppletion according to tense, because of the different stem between *I go* (present tense) and *I went* (past tense); it does not have suppletion according to aspect, because the stem is the same in *I go* (continuous aspect) and *I am going* (progressive aspect). The possible values for this feature are

1. Suppletion according to tense
2. Suppletion according to aspect
3. Suppletion in both tense and aspect
4. No suppletion in tense or aspect

In such cases, it is straightforward to recode this feature with two different features (Suppletion according to tense, Suppletion according to aspect).

For the number of genders (feature 30), we assume that a language with n genders can evolve in three possible ways: it can create a gender ($n \rightarrow n + 1$ genders), delete a gender ($n \rightarrow n - 1$ genders), or delete all genders at the same time ($n \rightarrow 0$ genders). The evolution of the feature is therefore modeled as a stepwise function, with the additional possibility of a catastrophic event. Such a catastrophic event occurred in the transition from Old English to Modern English: Old English has three genders, and Modern English has none (except

for the distinction *he/she/it*), but at no point did English have two genders. On the other hand, the three genders in Latin became two in French, Italian and Spanish ($n \rightarrow n - 1$).

A similar process is assumed for all the features which show some form of gradation, with one or two unusual states (for example, number of cases).

The catastrophic events should be less frequent than the addition or subtraction of a single gender. For the maximum parsimony calculations, that is rendered by penalizing the corresponding site; for the Bayesian implementation, we use an inverse-gamma prior, the tail of which is long enough to take this into account. The naive distance allowed rate heterogeneity between features, but it did not allow rate heterogeneity within a feature; this is now included in the model.

4.1 Results

4.1.1 Languages from different families

We used this refined distance on our set of 29 languages from different families and 106 features, to test the ability of the method to classify languages in correct families. The results were similar to those in section 3.2.2: the same groupings were found, but this time with a larger gap between the true families and the groupings which do not correspond to families. The same phenomenon was observed when using the improved distance with the 52 slowly-evolving features from section 3.2.3. While this distance does not help find new families, it does make identifying them easier.

4.1.2 Languages of Eurasia

We applied the improved distance and the 52 slowly-evolving features to 19 languages of Eurasia, in order to test several proposed languages families. The relationships between some of the languages in Fig. 6 are uncontroversial (Indo-European family, Uralic family, Nakh-Daghestanian family), but we wanted to test possible groupings such as:

- Altaic family: Turkish, Khalkha, Evenki, possibly Japanese and Korean
- Indo-Uralic superfamily: Indo-European and Uralic languages
- Uralo-Siberian superfamily: Uralic languages and Yukaghir ³
- Ural-Altaic superfamily: Uralic and some or all Altaic languages

These groupings have been proposed by some linguists on the basis on lexical similarities. None of them are accepted by the entire community.

The sample of languages was chosen so that supposedly related languages are geographically distant, making geographical contact unlikely. When languages which are in contact share features, we cannot differentiate genetical relationship from similarity due to contact; when such resemblances are found, we therefore interpret them as evidence for Sprachbunds rather than for genealogical relationship.

³Some proponents of this hypothesis add the Eskimo-Aleut family, found from Alaska to Greenland. We did not test this part of the proposed family.

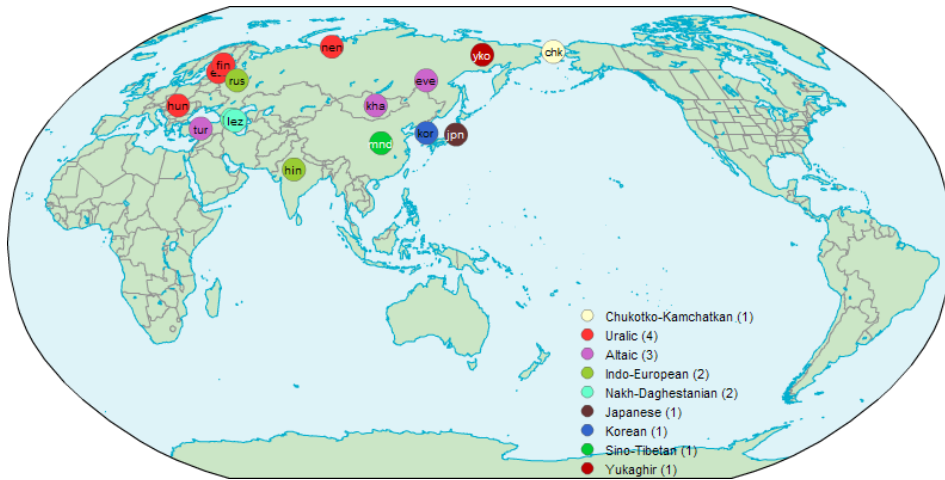


Figure 6: *Languages of Eurasia, shown by family, made using the electronic version of Dryer et al. (2005).. The Altaic grouping is controversial.*
Chk, Chukchi; Est, Estonian; Eve, Evenki; Fin, Finnish; Hin, Hindi; Hun, Hungarian; Hzb, Hunzib (hidden by Lez); Jpn, Japanese; Kor, Korean; Kha, Khalkha; Lez, Lezgian; Mnd, Mandarin; Nen, Nenets; Rus, Russian; Tur, Turkish; Yko, Yukaghir (Kolyma dialect)

The results show a clear split: all uncontroversial groupings have a posterior probability over 0.95, and no other grouping has a posterior probability over 0.67 (with the exception of a Sprachbund between Siberian languages). The next most likely groupings were two Sprachbunds, between Siberian languages (Chukchi, Nenets, Yukaghir), and between languages around the China Sea (Japanese, Korean, Mandarin; posterior probability: 0.67).

The Indo-Uralic superfamily and a superfamily including Nakh-Daghestanian languages, Khalkha, and the China sea Sprachbund were the only other two groupings with reasonably high probability (0.65). Notably, no evidence was found for the Altaic family, even in its reduced form (Evenki, Khalkha, Turkish): Turkish was found either to be a language isolate or to form an isolated group with Evenki (posterior probability 0.38); Khalkha was found to be either an isolate or a member of the China Sea Sprachbund (areal contact cannot be outruled here).

5 Conclusions

We showed that syntactical features can, to some extent, be used to build phylogenies, which correspond quite closely to those found by other methods. While the maximum parsimony method could find some relationships between languages, it was outperformed by maximum likelihood estimation of the tree, which was better both at clustering genera and at showing relationships within a genera. The results were robust to changes in the prior distribution.

Not all syntactical features evolve in the same way. Some are better retained from ancestor to daughter language, while others evolve more rapidly and/or

are more easily borrowed from neighbouring languages. The way the data is coded in the WALS database can also make relationships between languages less obvious.

The trees obtained are all unrooted. Rooting the trees could be done using some ancient language such as Hittite, which is an Indo-European language that diverged very early in the history of the Indo-European family.

This is only a first attempt to deal with some of the issues that arise when using syntactical (or other) features. It is certainly possible to come up with better models of how each feature evolves; an analysis of the correlation between features is also needed. This would highly reduce the number of usable features, but these independent features could then be used in conjunction with phonological or lexical features, as suggested in Dunn et al. (2005).

Nonetheless, we managed to reconstruct most uncontroversial language groupings, both genera and families. Our application to some more controversial groupings shows little or no evidence for them. One of the main issues is differentiation between borrowing and genetic transmission, or between Sprachbunds and families or genera.

As such, the results cannot be used for dating the ancestors, because there is too much variability in the branch lengths. Theoretically, however, it should be possible to date the ancestors, since a large number of calibration points are available (there is detailed data on the splits of Romance languages from Latin, or of Slavic languages from Old Church Slavonic, for example), which would provide more stability than the single calibration point often used in molecular genetics (the human/chimp split) (Atkinson et al., 2005). It would also be worth investigating whether there is a "syntactical clock", similar to the molecular clock.

6 Acknowledgements

I gratefully acknowledge Prof. Jotun Hein and Dr. Thomas Mailund for their helpful ideas, comments and suggestions.

Most of the data used comes from the WALS database (Dryer et al., 2005). Several people helped me collect additional data, or gain more insight into the data: Riyaz Baht (Kashmiri), Eleni Giannoulatou (Greek), Karine M. (Armenian), Thomas Mailund (Danish), Florence Paquet (Dutch).

Data was also taken from the following sources: Marchant (2004); Nichols (1999); Comrie et al. (1987).

Appendix

A Description of the languages

The following languages were used in this study. All information listed here is from Dryer et al. (2005) and Gordon (2005).

Language	Classification	Region	Speakers	Note
Albanian	Indo-European	Balkans	6 million	No close cousins; Balkan Sprachbund
Apuriña	Arawakan	Amazonia	2,000	
Armenian	Indo-European	Armenia	7 million	
Avestan	Indo-European, Indo-Iranian, Iranian	Persia		Extinct around 800 B.C.
Baluchi	Indo-European, Indo-Iranian, Western Iranian	Balochistan (Pakistan)	8 million	
Bawm	Sino-Tibetan, Tibeto-Burman	Indian subcontinent	14,000	
Belarusian	Indo-European, Slavic	Belarus	8 million	
Breton	Indo-European, Celtic	Brittany (France)	300,000	
Bulgarian	Indo-European, Slavic	Bulgaria	10 million	Balkan Sprachbund
Burmese	Sino-Tibetan, Tibeto-Burman	South-East Asia	42 million	
Canela-Krahô	Macro-Ge	North-East Brazil	2,500	
Chukchi	Chukotko-Kamchatkan	Chukotka (Russia)	10,000	
Czech	Indo-European, Slavic	Czech Republic	12 million	
Danish	Indo-European, Germanic, North Germanic	Denmark	5.5 million	
Dutch	Indo-European, Germanic, West Germanic	Belgium, Netherlands	22 million	
English	Indo-European, Germanic, West Germanic	England	400 million	
Estonian	Uralic, Finno-Ugric, Finnic	Estonia	1.1 million	
Evenki	(Altaic), Tungusic	Siberia	7,500	
Finnish	Uralic, Finno-Ugric, Finnic	Finland	6 million	
French	Indo-European, Romance	France	270 million	
German	Indo-European, Germanic, West Germanic	Germany	120 million	
Garô	Sino-Tibetan, Tibeto-Burman	India, Bangladesh	1 million	
Gooniyandi	Australian, Bunuban	Western Australia	100	Nearly extinct
Greek (Ancient)	Indo-European, Greek	Greece		Spoken from 1100 B.C. to 600 A.D.
Greek (Modern)	Indo-European, Greek	Greece	15 million	No close cousins; Balkan Sprachbund
Greenlandic	Eskimo-Aleut	Greenland	54,000	aka Kalaallisut
Hausa	Afro-Asiatic, West Chadic	Nigeria	40 million	
Hindi	Indo-European, Indo-Iranian, Indo-Aryan	India	800 million	
Hungarian	Uralic, Finno-Ugric, Ugric	Hungary	13 million	
Hunzib	Nakh-Daghestanian	North Caucasus	2,000	
Icelandic	Indo-European, Germanic, West Germanic	Iceland	300,000	
Igbo	Niger-Congo, Benue-Congo, Igboid	Nigeria	18 million	
Ingush	Nakh-Daghestanian	Ingushethia, Chechnya	230,000	
Irish	Indo-European, Celtic	Ireland	250,000	
Italian	Indo-European, Romance	Italy	70 million	
Japanese	Japanese	Japan	130 million	
Ju'hoan	Khoisan	Namibia, Botswana	30,000	aka !Xu, !Kung or Qxü
Kashmiri	Indo-European, Indo-Iranian, Dardic	Kashmir (India, Pakistan)	4.6 million	
Kera	Afro-Asiatic, East Chadic	Chad	50,000	
Khalkha	(Altaic), Mongolic	Mongolia	2.3 million	
Khoekhoe	Khoisan	Southern Africa	250,000	aka Nàmá
Korean	Korean	Korea	78 million	
Latin	Indo-European, Italic	Europe		extinct
Latvian	Indo-European, Baltic	Latvia	3.2 million	
Lezgian	Nakh-Daghestanian	Central Caucasus	450,000	
Lithuanian	Indo-European, Baltic	Lithuania	4 million	
Luvale	Niger-Congo, Benue-Congo, Bantoid	Angola, Zambia	670,000	
Macedonian	Indo-European, Slavic	Macedonia	2 million	Balkan Sprachbund

Maithili	Indo-European, Indo-Iranian, Indo-Aryan,	Nepal, Bihar (India)	24 million	
Mandarin	Sino-Tibetan, Chinese	China	867 million	
Maranungku	Australian	Northern Australia	15-20 (1983)	Nearly extinct
Marathi	Indo-European, Indo-Iranian, Indo-Aryan	Maharashtra (India)	90 million	
Maricopa	Hokan	Native American	181 (1990)	Nearly extinct; all speakers are older adults
Matuthunira	Australian	Western Australia	5 (1981)	Nearly extinct
Meithei	Sino-Tibetan, Tibeto-Burman	India	1,200,000	
Nenets	Uralic, samoyedic	Siberia		
Nkore-Kiga	Niger-Congo, Benue-Congo, Bantoid	Uganda	1,400,000	aka Chiga
Nunggubuyu	Australian, Gunwinyguan	Northern Australia	700	
Oromo	Afro-Asiatic, Eastern Suchitic	Ethiopia, Kenya	25 million	Data for the Harar dialect of Oromo
Ossetic	Indo-European, Indo-Iranian, Eastern Iranian	Ossetia (Russia, Georgia)	700,000	
Panjabi	Indo-European, Indo-Iranian, Eastern Iranian	Panjab (India, Pakistan)	104 million	
Persian	Indo-European, Indo-Iranian, Iranian	Iran, Afghanistan, Tajikistan	110 million	aka Fārsi
Polish	Indo-European, Slavic	Poland	46 million	
Pomo	Hokan	Native American (California)	5 (1994)	Nearly extinct. Data for South-Eastern dialect.
Romanian	Indo-European, Romance	Romania	24 million	Balkan Sprachbund
Russian	Indo-European, Slavic	Asia, Eastern Europe	255 million	
Sango	Niger-Congo, Adamawa-Ubangian	Central African Republic	5 million	
Sanskrit	Indo-European, Indo-Iranian, Indo-Aryan	South Asia		Lithurgical language in Hinduism
Serbo-Croatian	Indo-European, Slavic	Balkans	17 million	Considered to be split into Bosnian, Serbian and Croatian
Slovak	Indo-European, Slavic	Slovakia	6 million	
Slovenian	Indo-European, Slavic	Slovenia	2.2 millio	
Somali	Afro-Asiatic, Eastern Cushitic	Somalia	20 million	
Spanish	Indo-European, Romance	Spain	410 million	
Swahili	Niger-Congo, Benue-Congo, Bantoid	Eastern Africa	50 million	
Tocharian A	Indo-European, Tocharian	Central Asia		Extinct; spoken between 600 and 800 A.D.
Turkish	(Altaic), Turkic	Turkey	75 million	
Ukrainian	Indo-European, Slavic	Ukraine	40 million	
Wardaman	Australian, Gunwinyguan	Northern Australia	50 (1983)	Nearly extinct
Welsh	Indo-European, Celtic	Wales	750,000	
Yukaghir	Yukaghir	Siberia	10-50	Data is for the Kolyma dialect
Yup'ik	Eskimo-Aleut	Alaska	10,000	

B Details of the features

This is a list of the syntactical features used in this study, and their state-space. A detailed description of the features is available in Dryer et al. (2005).

Morphology

- 20** Fusion of Selected Inflectional Formatives
 - 1. Exclusively concatenative
 - 2. Exclusively isolating
 - 3. Exclusively tonal
 - 4. Tonal/isolating
 - 5. Tonal/concatenative
 - 6. Ablaut/concatenative
 - 7. Isolating/concatenative
- 21** Exponence of Selected Inflectional Formatives
 - 1. Monoexponential case
 - 2. Case+number
 - 3. Case+referentiality
 - 4. Case+TAM (tense-aspect-mood)
 - 5. No case
- 22** Inflectional Synthesis of the Verb
 - 1. 0-1 categories per word
 - 2. 2-3 categories per word
 - 3. 4-5 categories per word
 - 4. 6-7 categories per word
 - 5. 8-9 categories per word
 - 6. 10-11 categories per word
 - 7. 12-13 categories per word
- 23** Locus of Marking in the Clause
 - 1. P is head-marked
 - 2. P is dependent-marked
 - 3. P is double-marked
 - 4. P has no marking
 - 5. Other types
- 24** Locus of Marking in Possessive Noun Phrases
 - 1. Possessor is head-marked
 - 2. Possessor is dependent-marked
 - 3. Possessor is double-marked
 - 4. Possessor has no marking
 - 5. Other types
- 25** Locus of Marking: Whole-Language Typology
 - 1. Consistently head-marking
 - 2. Consistently dependent-marking
 - 3. Consistently double-marking
 - 4. Consistently zero-marking
 - 5. Inconsistent marking or other type
- 26** Prefixing vs. Suffixing in Inflectional Morphology
 - 1. Little or no inflectional morphology
 - 2. Predominantly suffixing
 - 3. Moderate preference for suffixing
 - 4. Approximately equal amounts of suffixing and prefixing
 - 5. Moderate preference for prefixing
 - 6. Predominantly prefixing
- 27** Reduplication
 - 1. Productive full and partial reduplication
 - 2. Full reduplication only
 - 3. No productive reduplication
- 28** Case Syncretism
 - 1. Inflectional case marking is absent or minimal
 - 2. Inflectional case marking is syncretic for core cases only
 - 3. Inflectional case marking is syncretic for core and non-core cases
 - 4. Inflectional case marking is never syncretic
- 29** Syncretism in Verbal Person/Number Marking
 - 1. No subject person/number marking
 - 2. Subject person/number is syncretic
 - 3. Subject person/number is never syncretic

Nominal Categories

- 30** Number of Genders
 - 1. None
 - 2. Two
 - 3. Three
 - 4. Four
 - 5. Five or more
- 31** Sex-based and Non-sex-based Gender Systems
 - 1. No gender system
 - 2. Sex-based gender system
 - 3. Non-sex-based gender system
- 32** Systems of Gender Assignment
 - 1. No gender system
 - 2. Semantic assignment
 - 3. Semantic and formal assignment
- 33** Coding of Nominal Plurality
 - 1. Plural prefix
 - 2. Plural suffix
 - 3. Plural stem change
 - 4. Plural tone
 - 5. Plural by complete reduplication of stem
 - 6. Morphological plural with no method primary
 - 7. Plural word
 - 8. Plural clitic
 - 9. No plural
- 34** Occurrence of Nominal Plurality
 - 1. No nominal plural
 - 2. Plural only in human nouns, optional
 - 3. Plural only in human nouns, obligatory
 - 4. Plural in all nouns, always optional
 - 5. Plural in all nouns, optional in inanimates
 - 6. Plural in all nouns, always obligatory
- 35** Plurality in Independent Personal Pronouns
 - 1. No independent subject pronouns
 - 2. Number-indifferent pronouns
 - 3. Person-number affixes
 - 4. Person-number stem
 - 5. Person-number stem with a pronominal plural affix
 - 6. Person-number stem with a nominal plural affix
 - 7. Person stem with a pronominal plural affix
 - 8. Person stem with a nominal plural affix
- 36** The Associative Plural
 - 1. Associative plural marker also used for additive plurals
 - 2. Special bound associative plural marker
 - 3. Special non-bound associative plural marker
 - 4. Associative plural absent
- 37** Definite Articles
 - 1. Definite word distinct from demonstrative
 - 2. Demonstrative word used as marker of definiteness
 - 3. Definite affix on noun
 - 4. No definite article but indefinite article
 - 5. Neither definite nor indefinite article
- 38** Indefinite Articles
 - 1. Indefinite word distinct from numeral for 'one'
 - 2. Numeral for 'one' is used as indefinite article
 - 3. Indefinite affix on noun
 - 4. No indefinite article but definite article
 - 5. Neither indefinite nor definite article

- 39 Inclusive/Exclusive Distinction in Independent Pronouns
 - 1. No grammaticalised marking at all
 - 2. 'We' and 'I' identical
 - 3. No inclusive/exclusive opposition
 - 4. Only inclusive differentiated
 - 5. Inclusive and exclusive differentiated
- 40 Inclusive/Exclusive Distinction in Verbal Inflection
 - 1. No inclusive/ exclusive opposition
 - 2. Inclusive and exclusive differentiated
- 41 Distance Contrasts in Demonstratives
 - 1. No distance contrast
 - 2. Two-way contrast
 - 3. Three-way contrast
 - 4. Four-way contrast
 - 5. Five (or more)-way contrast
- 42 Pronominal and Adnominal Demonstratives
 - 1. same forms
 - 2. different stems
 - 3. different inflectional features
- 43 Third Person Pronouns and Demonstratives
 - 1. Unrelated
 - 2. Related for all demonstratives
 - 3. Related to remote demonstratives
 - 4. Related to non-remote demonstratives
 - 5. Related by gender markers
 - 6. Related for nonhuman reference
- 44 Gender Distinctions in Independent Personal Pronouns
 - 1. Gender distinctions in 3rd person plus 1st and/or 2nd person
 - 2. Gender distinctions in 3rd person only, but in both singular and nonsingular
 - 3. Gender distinctions in 3rd person singular only
 - 4. Gender distinctions in 1st or 2nd person but not 3rd
 - 5. Gender distinctions in 3rd person non-singular only
 - 6. No gender distinctions
- 45 Politeness Distinctions in Pronouns Second person pronouns
 - 1. encode no politeness distinction
 - 2. encode a binary politeness distinction
 - 3. encode multiple politeness distinctions
 - 4. are dominantly avoided for politeness reasons
- 46 Indefinite Pronouns
 - 1. Interrogative-based indefinites
 - 2. Generic-noun-based indefinites
 - 3. Special indefinites
 - 4. Mixed indefinites
 - 5. Existential construction
- 47 Intensifiers and Reflexive Pronouns
 - 1. Intensifiers and reflexive pronouns are formally identical
 - 2. Intensifiers and reflexive pronouns are formally differentiated
- 48 Person Marking on Adpositions
 - 1. No adpositions
 - 2. Adpositions without person marking
 - 3. Person marking for pronouns only
 - 4. Person marking for pronouns and nouns
- 49' Morphological case-marking
 - 0 No morphological case-marking
 - 1 Presence of morphological case-marking
- 49 Number of Cases
 - 1. No morphological case-marking

2. 2 case categories
 3. 3 case categories
 4. 4 case categories
 5. 5 case categories
 6. 6-7 case categories
 7. 8-9 case categories
 8. 10 or more case categories
 9. Exclusively borderline morphological case-marking
- 50 Assymetrical Case-marking**
1. No morphological case-marking
 2. Symmetrical case-marking
 3. Additive-quantitatively asymmetrical case-marking
 4. Subtractive-quantitatively asymmetrical case-marking
 5. Qualitatively asymmetrical casemarking
 6. Syncretism in relevant NP-types
- 51 Position of Case Affixes**
1. Case suffixes
 2. Case prefixes
 3. Case coded by tone
 4. Case coded by changes within noun stem
 5. Mixed morphological case strategies with none primary
 6. Postpositional clitics
 7. Prepositional clitics
 8. Inpositional clitics
 9. Neither case affixes nor adpositional clitics
- 52 Comitatives and Instrumentals**
1. Identity
 2. Differentiation
 3. Mixed
- 53 Ordinal Numerals**
1. Zero: Ordinal numerals do not exist
 2. One: No distinction of cardinal and ordinal numerals
 3. First: Cardinal and ordinal numerals are identical except for one and first
 4. One-th: Ordinal numerals are derived from cardinal numerals
 5. First/One-th: All ordinal numerals are derived from cardinal numerals with two alternatives for first, one of which is morphologically independent of one
 6. Two-th: Ordinal numerals from two upwards are derived from cardinal numerals, first is suppletive
 7. Second: First and a small set of consecutive higher ordinal numerals are suppletive
 8. Variou-th: Other solutions
- 54 Distributive Numerals**
1. No distributive numerals
 2. Marked by reduplication
 3. Marked by prefix
 4. Marked by suffix
 5. Marked by preceding word
 6. Marked by following word
 7. Marked by mixed or other strategies
- 55 Numeral Classifiers**
1. Numeral classifiers are absent
 2. Numeral classifiers are optional
 3. Numeral classifiers are obligatory
- 56 Conjunctions and Universal Quantifiers**
1. Formally different
 2. Formally similar, not involving interrogative expression
 3. Formally similar, involving interrogative expression
- 57 Position of Pronominal Possessive Affixes**
1. Possessive prefixes
 2. Possessive suffixes
 3. Both possessive prefixes and possessive suffixes, with neither primary
 4. No possessive affixes

Nominal Syntax

58 Obligatory Possessive Inflection

1. Obligatory possessed nouns exist
2. No obligatory possessed nouns

59 Possessive Classification

1. No possessive classification
2. Two classes
3. Three to five classes
4. More than five classes

60 Genitives, Adjectives and Relative Clauses

1. Weakly differentiated
2. Moderately differentiated, with genitives and adjectives collapsed
3. Moderately differentiated, with genitives and relative clauses collapsed
4. Moderately differentiated, with adjectives and relative clauses collapsed
5. Moderately differentiated; other
6. Highly differentiated

61 Adjectives without Nouns

1. Adjective may not occur without noun
2. Adjective may occur without noun, and without marking
3. Adjective may occur without noun, obligatorily marked by prefix
4. Adjective may occur without noun, obligatorily marked by suffix
5. Adjective may occur without noun, obligatorily marked by preceding word
6. Adjective may occur without noun, obligatorily marked by following word
7. Adjective may occur without noun, obligatorily marked by mixed or other strategies

62 Action Nominal Constructions

1. Sentential: dependent-marking of the finite clause is retained for S, A and P
2. Possessive-Accusative: S/A treated as possessors, P retains sentential marking
3. Ergative-Possessive: S/P treated as possessors, A treated differently
4. Double-Possessive: All major arguments treated as possessors
5. Other: Minor patterns
6. Mixed: Several patterns in the same language
7. Not both A and P in the same construction
8. No action nominals

63 Noun Phrase Conjunction

1. AND-languages: 'and' and 'with' are not identical
2. WITH-languages: 'and' and 'with' are identical

64 Nominal and Verbal Conjunction

1. Nominal and verbal conjunction are largely identical
2. Nominal and verbal conjunction are different
3. Nominal and verbal conjunction are primarily expressed by juxtaposition

Verbal Categories

65 Perfective/Imperfective Aspect

1. Grammatical marking of perfective/imperfective distinction
2. No grammatical marking of perfective/imperfective distinction

66 The Past Tense

1. Past/non-past distinction marked; no remoteness distinction
2. Past/non-past distinction marked; 2–3 degrees of remoteness distinguished
3. Past/non-past distinction marked; at least 4 degrees of remoteness distinguished
4. No grammatical marking of past/nonpast distinction

67 The Future Tense

1. Inflectional marking of future/nonfuture distinction
2. No inflectional marking of future/non-future distinction

68 The Perfect

1. Perfect of the have-type (derived from a possessive construction)
2. Perfect derived from word meaning finish or already

- 3. Other perfect
 - 4. No perfect
- 69** Position of Tense-Aspect Affixes
- 1. Tense-aspect prefixes
 - 2. Tense-aspect suffixes
 - 3. Tense-aspect tone
 - 4. Combination of tense-aspect strategies with none primary
 - 5. No tense-aspect inflection
- 70** The Morphological Imperative
- 1. The language has morphologically dedicated second singular as well as second plural imperatives
 - 2. The language has morphologically dedicated second singular imperatives but no morphologically dedicated second plural imperatives
 - 3. The language has morphologically dedicated second plural imperatives but no morphologically dedicated second singular imperatives
 - 4. The language has morphologically dedicated second person imperatives that do not distinguish between singular and plural
 - 5. The language has no morphologically dedicated second person imperatives at all
- 71** The Prohibitive
- 1. The prohibitive uses the verbal construction of the second singular imperative and a sentential negative strategy found in (indicative) declaratives
 - 2. The prohibitive uses the verbal construction of the second singular imperative and a sentential negative strategy not found in (indicative) declaratives
 - 3. The prohibitive uses a verbal construction other than the second singular imperative and a sentential negative strategy found in (indicative) declaratives
 - 4. The prohibitive uses a verbal construction other than the second singular imperative and a sentential negative strategy not found in (indicative) declaratives
- 72** Imperative-Hortative Systems
- 1. The language has a maximal system, but not a minimal one
 - 2. The language has a minimal system, but not a maximal one
 - 3. The language has both a maximal and a minimal system
 - 4. The language has neither a maximal nor a minimal system
- 73** The Optative
- 1. inflectional optative present
 - 2. inflectional optative absent
- 74** Situational Possibility
- 1. The language can express situational possibility with affixes on verbs
 - 2. The language does not express situational possibility with affixes on verbs, but with verbal constructions
 - 3. The language does not express situational possibility with affixes on verbs or with verbal constructions, but with other kinds of markers
- 75** Epistemic Possibility
- 1. The language can express epistemic possibility with verbal constructions
 - 2. The language does not express epistemic possibility with verbal constructions, but with affixes on verbs
 - 3. The language does not express epistemic possibility with verbal constructions or with affixes on verbs, but with other kinds of markers
- 76** Overlap between Situation and Epistemic Modal Marking
- 1. The language has markers that can code both situational and epistemic modality, both for possibility and for necessity
 - 2. The language has markers that can code both situational and epistemic modality, but only for possibility or only for necessity
 - 3. The language has no markers that can code both situational and epistemic modality
- 77** Semantic Distinctions of Evidentiality
- 1. No grammatical evidentials
 - 2. Only indirect evidentials
 - 3. Both direct and indirect evidentials
- 78** Coding of Evidentiality
- 1. No grammatical evidentials
 - 2. Verbal affix or clitic
 - 3. Part of the tense system
 - 4. Separate particle
 - 5. Modal morpheme

- 6. Mixed systems
- 79** Suppletion According to Tense and Aspect
 - 1. Suppletion according to tense
 - 2. Suppletion according to aspect
 - 3. Suppletion in both tense and aspect
 - 4. No suppletion in tense or aspect
- 80** Verbal Number and Suppletion
 - 1. No singular-(dual)-plural pairs/triples in the reference material
 - 2. Singular-plural pairs, no suppletion
 - 3. Singular-plural pairs, suppletion
 - 4. Singular-dual-plural triples, no suppletion
 - 5. Singular-dual-plural triples, suppletion

Word Order

- 81** Order of Subject, Object and Verb
 - 1. Subject-Object-Verb (SOV)
 - 2. Subject-Verb-Object (SVO)
 - 3. Verb-Subject-Object (VSO)
 - 4. Verb-Object-Subject (VOS)
 - 5. Object-Verb-Subject (OVS)
 - 6. Object-Subject-Verb (OSV)
 - 7. Lacking a dominant word order
- 82** Order of Subject and Verb
 - 1. Subject precedes verb (SV)
 - 2. Subject follows verb (VS)
 - 3. Both orders with neither order dominant
- 83** Order of Object and Verb
 - 1. Object precedes verb (OV)
 - 2. Object follows verb (VO)
 - 3. Both orders with neither order dominant
- 84** Order of Object, Oblique and Verb
 - 1. Verb-object-oblique order (VOX)
 - 2. Oblique-verb-object order (XVO)
 - 3. Oblique-object-verb order (XOV)
 - 4. Object-oblique-verb order (OXV)
 - 5. Object-verb-oblique order (OVX)
 - 6. More than one order with none dominant
- 85** Order of Adposition and Noun Phrases
 - 1. Postpositions
 - 2. Prepositions
 - 3. Inpositions
 - 4. More than one adposition type with none dominant
 - 5. No adpositions
- 86** Order of Genitive and Noun
 - 1. Genitive-noun (GenN)
 - 2. Noun-genitive (NGen)
 - 3. Both orders occur with neither order dominant
- 87** Order of Adjective and Noun
 - 1. Modifying adjectives precedes noun (AdjN)
 - 2. Modifying adjectives follows noun (NAdj)
 - 3. Both orders of noun and modifying adjective occur, with neither dominant
 - 4. Adjectives do not modify nouns, occurring as predicates in internally headed relative clauses
- 88** Order of Demonstrative and Noun
 - 1. Demonstrative word precedes noun (DemN)
 - 2. Demonstrative word follows noun (NDem)
 - 3. Demonstrative prefix on noun
 - 4. Demonstrative suffix on noun

5. Demonstrative simultaneously before and after noun
 6. Two or more of above types with none dominant
- 89** Order of Numeral and Noun
1. Numeral precedes noun (NumN)
 2. Numeral follows noun (NNum)
 3. Both orders of numeral and noun with neither order dominant
 4. Numeral only modifies verb
- 90** Order of Relative Clause and Noun
1. Relative clause follows noun (NRel)
 2. Relative clause precedes noun (RelN)
 3. Internally-headed relative clause
 4. Correlative relative clause
 5. Adjoined relative clause
 6. Double-headed relative clause
 7. Mixed types of relative clause with none dominant
- 91** Order of Degree Word and Adjective
1. Degree word precedes adjective (DegAdj)
 2. Degree word follows adjective (AdjDeg)
 3. Both orders occur with neither order dominant
- 92** Position of Polar Question Particles
1. Question particle at beginning of sentence
 2. Question particle at end of sentence
 3. Question particle in second position in sentence
 4. Question particle with other position
 5. Question particle in either of two positions
 6. No question particle
- 93** Position of Interrogative Phrases in Content Questions
1. Interrogative phrases obligatorily initial
 2. Interrogative phrases not obligatorily initial
 3. Mixed, some interrogative phrases obligatorily initial, some not
- 94** Order of Adverbial Subordinator and Clause
1. Adverbial subordinators which are separate words and which appear at the beginning of the subordinate clause
 2. Adverbial subordinators which are separate words and which appear at the end of the subordinate clause
 3. Clause-internal adverbial subordinators
 4. Suffixal adverbial subordinators
 5. More than one type of adverbial subordinator with none dominant

Simple Clauses

- 98** Alignment of Case Marking of Full Noun Phrases
1. Neutral
 2. Nominative-accusative (standard)
 3. Nominative-accusative (marked nominative)
 4. Ergative-absolutive
 5. Tripartite
 6. Active-inactive
- 99** Alignment of Case Marking of Pronouns
1. Neutral
 2. Nominative-accusative (standard)
 3. Nominative-accusative (marked nominative)
 4. Ergative-absolutive
 5. Tripartite
 6. Active-inactive
 7. None
- 100** Alignment of Verbal Person Marking
1. Neutral alignment (no verbal person marking)
 2. Accusative alignment

3. Ergative alignment
 4. Active alignment
 5. Hierarchical alignment
 6. Split alignment
- 101 Expression of Pronominal Subject**
1. Pronominal subjects are expressed by pronouns in subject position that are normally if not obligatorily present
 2. Pronominal subjects are expressed by affixes on verbs
 3. Pronominal subjects are expressed by clitics with variable host
 4. Pronominal subjects are expressed by subject pronouns that occur in a different syntactic position from nominal subjects
 5. Pronominal subjects are expressed only by pronouns in subject position, but these pronouns are often left out
 6. More than one of the above types with none dominant
- 102 Verbal Person Marking**
1. No person marking of any argument
 2. Person marking of only the A argument
 3. Person marking of only the P argument
 4. Person marking of the A or P argument
 5. Person marking of both the A and P arguments
- 103 Third Person Zero of Verbal Person Marking**
1. No person marking of the S
 2. No zero realization of third person S forms
 3. Zero realization of some third person singular S forms
 4. Zero realization of all third person singular S forms
 5. Zero realization of all third person S forms/No third person S forms
 6. Zero realization only of third person nonsingular S
- 104 Order of Person Markers on the Verb**
1. A and P do not, or do not both, occur on the verb
 2. A precedes P
 3. P precedes A
 4. Both orders of A and P occur
 5. A and P are fused
- 105 Ditransitive Constructions: The Verb 'Give'**
1. Indirect-object construction
 2. Double-object construction
 3. Secondary-object construction
 4. Mixed
- 106 Reciprocal Constructions**
1. There are no non-iconic reciprocal constructions
 2. All reciprocal constructions are formally distinct from reflexive constructions
 3. There are both reflexive and nonreflexive reciprocal constructions
 4. The reciprocal and reflexive constructions are formally identical
- 107 Passive Constructions**
1. There is a passive construction
 2. There is no passive construction
- 108 Antipassive Constructions**
1. Antipassive with patient-like argument left implicit
 2. Antipassive with patient-like argument expressed as oblique complement
 3. No antipassive
- 109 Applicative Constructions**
1. Benefactive object only; both bases
 2. Benefactive object only; transitive base only
 3. Benefactive and other; both bases
 4. Benefactive and other; transitive base only
 5. Non-benefactive object only; both bases
 6. Non-benefactive object only; transitive base only
 7. Non-benefactive object only; intransitive base only
 8. No applicative construction

- 110 Periphrastic Causative Constructions
 - 1. Sequential type but no purposive type
 - 2. Purposive type but no sequential type
 - 3. Both sequential type and purposive type
- 111 Nonperiphrastic Causative Constructions
 - 1. No morphological type or compound type
 - 2. Morphological type but no compound type
 - 3. Compound type but no morphological type
 - 4. Both morphological type and compound type
- 112 Negative Morphemes
 - 1. Negative affix
 - 2. Negative particle
 - 3. Negative auxiliary verb
 - 4. Negative word, unclear if verb or particle
 - 5. Variation between negative word and affix
 - 6. Double negation
- 113 Symmetric and Asymmetric Standard Negation
 - 1. Symmetric standard negation only: Type Sym
 - 2. Asymmetric standard negation only: Type Asy
 - 3. Symmetric and asymmetric standard negation: Type SymAsy
- 114 Subtypes of Asymmetric Standard Negation
 - 1. in finiteness: Subtype A/Fin
 - 2. in reality status: Subtype A/NonReal
 - 3. in other grammatical categories: Subtype A/Cat
 - 4. in finiteness and reality status: Subtypes A/Fin and A/NonReal
 - 5. in finiteness and other grammatical categories: Subtypes A/Fin and A/Cat
 - 6. in reality status and other grammatical categories: Subtypes A/NonReal and A/Cat
 - 7. non-assignable (no asymmetry found)
- 115 Negative Indefinite Pronouns and Predicate Negation
 - 1. Negative indefinites co-occur with predicate negation
 - 2. Negative indefinites preclude predicate negation
 - 3. Negative indefinites show mixed behaviour
 - 4. Negative existential construction
- 116 Polar Questions
 - 1. Question particle
 - 2. Interrogative verb morphology
 - 3. Question particle and interrogative verb morphology
 - 4. Interrogative word order
 - 5. Absence of declarative morphemes
 - 6. Interrogative intonation only
 - 7. No interrogative-declarative distinction
- 117 Predicative Possession
 - 1. Locational Possessive
 - 2. Genitive Possessive
 - 3. Topic Possessive
 - 4. Conjunctive Possessive
 - 5. Have-Possessive
- 118 Predicative Adjectives
 - 1. Predicative adjectives have verbal encoding
 - 2. Predicative adjectives have nonverbal encoding
 - 3. Predicative adjectives have mixed encoding
- 119 Nominal and Locational Predication
 - 1. Split (ie different) encoding of nominal and locational predication
 - 2. Shared (ie identical) encoding of nominal and locational predication
- 120 Zero Copula for Predicative Nominals
 - 1. Zero copula is impossible
 - 2. Zero copula is possible
- 121 Comparative Constructions
 - 1. Locational Comparative
 - 2. Exceed Comparative
 - 3. Conjoined Comparative
 - 4. Particle Comparative

Complex Sentence

122 Relativization on Subjects

1. Relative pronoun
2. Non-reduction
3. Pronoun-retention
4. Gap

123 Relativization on Obliques

1. Relative pronoun strategy
2. Nonreduction strategy
3. Pronoun retention strategy
4. Gap strategy
5. Not possible

124 'Want' Complement Subjects

1. The complement subject is left implicit
2. The complement subject is expressed overtly
3. Both construction types exist
4. 'Want' is expressed as a desiderative verbal affix
5. 'Want' is expressed as an uninflected desiderative particle

125 Purpose Clauses

1. Balanced
2. Balanced/ deranked
3. Deranked

126 'When' Clauses

1. Balanced
2. Balanced/ deranked
3. Deranked

127 Reason Clauses

1. Balanced
2. Balanced/ deranked
3. Deranked

128 Utterance Complement Clauses

1. Balanced
2. Balanced/ deranked
3. Deranked

References

- Q. Atkinson, G. Nicholls, D. Welch, and R. Gray. From words to dates: water into wine, mathemagic or phylogenetic inference? *Trans. Philol. Soc.*, 103(2): 193–219, 2005.
- B. Comrie. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press, 1989.
- B. Comrie et al. *The world's major languages*. Routledge, 1987.
- C. Darwin. *The descent of man and selection in relation to sex*. John Murray, 1871.
- J. Diamond and P. Bellwood. Farmers and Their Languages: The First Expansions. *Science*, 300(5619):597–603, 2003.
- M. Dryer. *Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase*. Oxford University Press, 2005.
- M. Dryer, M. Haspelmath, D. Gil, and B. Comrie. *World Atlas of Language Structures*, 2005.
- M. Dunn, A. Terrill, G. Reesink, R. Foley, and S. Levinson. Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science*, 309(5743):2072–2075, 2005.
- J. Felsenstein. **PARS** - discrete character parsimony. Available online at <http://evolution.genetics.washington.edu/phylip/doc/pars.html>, 2004.
- J. Felsenstein. **PHYLIP** (phylogeny inference package) version 3.65. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005.
- W. Gilks and G. Roberts. Strategies for improving MCMC. *Markov Chain Monte Carlo in Practice*, pages 89–114, 1996.
- R. G. Gordon, editor. *Ethnologue: Languages of the World, Fifteenth edition*. Tex.: SIL International, 2005. Online version: <http://www.ethnologue.com/>.
- R. Gray and Q. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–9, 2003.
- R. Gray and F. Jordan. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405(6790):1052–1055, 2000.
- B. D. Joseph. The Indo-European Family - The Linguistic Evidence. In A.-P. Christides, editor, *History of the Greek Language from the beginnings up to later antiquity*, pages 128–134. Thessaloniki: Centre for the Greek Language, 2001.
- C. F. Justus. Indo-european documentation center - semantic fields. Available online at <http://www.utexas.edu/cola/centers/lrc/iedocctr/ie-ling/ie-sem/ie-sem.html>, 2006.

- C. Marchant. *Fundamentals of Modern Belarusian*, 2004.
- Y. Matras. How predictable is contact-induced change in grammar? In C. Renfrew, A. McMahon, and L. Trask, editors, *Time Depth in Historical Linguistics*, pages 563–583. McDonald Institute for Archaeological Research, 2000.
- A. McMahon and R. McMahon. *Language Classification by Numbers*. Oxford Linguistics, 2005.
- J. Nichols. *Linguistic Diversity in Space and Time*. University of Chicago Press, 1999.
- R. Page. TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, 12(4):357–358, 1996.
- D. Ringe, T. Warnow, and S. Evans. Polymorphic characters in Indo-European languages. *Languages and Genes*, September 2006. Conference talk.
- M. Rosenfelder. How likely are chance resemblances between languages? Available online at <http://www.zompist.com/chance.htm>, 2002.
- M. Rosenfelder. Proto-World and the Language Instinct. Available online at <http://www.zompist.com/langorg.htm>, 2006.
- M. Stoneking. Disentangling Genes, Geography, and Language. *Languages and Genes*, September 2006. Conference talk.
- S. Thomason and T. Kaufman. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, 1988.
- R. Trask. *Historical Linguistics*. Arnold, 1996.
- S. Wichmann and A. Saunders. How to use typological databases in historical linguistic research. Revised submission under review for *Diachronica*, 2006ms.