Using Ancestral State Reconstruction Methods for Onomasiological Reconstruction in Multilingual Word Lists

Draft, accepted for publication in Language Dynamics and Change

Gerhard Jäger and Johann-Mattis List

June, 2017

Abstract

Current efforts in computational historical linguistics are predominantly concerned with phylogenetic inference. Methods for ancestral state reconstruction have been only sporadically applied. In contrast to phylogenetic algorithms, automatic reconstruction methods presuppose phylogenetic information in order to explain what has evolved when and where. Here we report a pilot study exploring how well automatic methods for ancestral state reconstruction perform in the task of onomasiological reconstruction in multilingual word lists, where algorithms are used to infer how the words evolved along a given phylogeny, and reconstruct which cognate classes were used to express a given meaning in the ancestral languages. Comparing three different methods, Maximum Parsimony, Minimal Lateral Networks, and Maximum Likelihood on three different test sets (Indo-European, Austronesian, Chinese) using binary and multi-state coding of the data as well as single and sampled phylogenies, we find that Maximum Likelihood largely outperformed the other methods. At the same time, however, the general performance was disappointingly low, ranging between 0.66 (Chinese) and 0.79 (Austronesian) for the F-Scores. A closer linguistic evaluation of the reconstructions proposed by the best method and the reconstructions given in the gold standards revealed that the majority of the cases where the algorithms failed can be attributed to problems of independent semantic shift (homoplasy), to morphological processes in lexical change, and to wrong reconstructions in the independently created test sets that we employed.

1 Introduction

Phylogenetic reconstruction methods are crucial for recent quantitative approaches in historical linguistics. While many scholars remain skeptical regarding the potential of methods for automatic sequence comparison, phylogenetic reconstruction, be it of networks using the popular SplitsTree software (Huson, 1998), or family trees, using distance-

(Sokal and Michener, 1958; Saitou and Nei, 1987) or character-based approaches (Edwards and Cavalli-Sforza, 1964; Fitch, 1971; Ronquist et al., 2012; Bouckaert et al., 2014), have entered the mainstream of historical linguistics. This is reflected in a multitude of publications and applications on different language families, be it Ainu (Lee and Hasegawa, 2013), Australian (Bowern and Atkinson, 2012), Semitic (Kitchen et al., 2009), or Chinese (Ben Hamed and Wang, 2006), and a growing interest in the implications of phylogenetic analyses for historical linguistics, as reflected in the heated debate about the dating of Indo-European (Gray and Atkinson, 2003; Atkinson and Gray, 2006; Bouckaert et al., 2014; Chang et al., 2015), and in recent attempts to search for deep genetic signals in the languages of the world (Pagel et al., 2013; Jäger, 2015).

Given the boom of quantitative approaches in the search for language trees and networks, it is surprising that methods which infer the ancestral states of linguistic characters, have been rarely applied and tested so far. While methods for phylogenetic reconstruction infer how related languages evolved into their current shape, methods for *ancestral state reconstruction* (ASR) use a given phylogeny to infer the previous appearance of the languages. This is illustrated in Figure 1 for the reconstruction of lexical conceptualization patterns (more on this specific kind of ancestral state reconstruction below). What is modeled as ancestral state in this context is open to the researcher's interest, ranging from the original pronunciation of words (Bouchard-Côté et al., 2013), the direction of sound change processes (Hruschka et al., 2015), the original conceptualization of concepts (List, 2016), or even linguistic and cultural aspects beyond the lexicon, such as ancestral color systems (Haynie and Bowern, 2016), numeral systems (Zhou and Bowern, 2015) or cultural patterns, such as matrilocality (Jordan et al., 2009). While methods for ancestral state reconstruction are commonly used in evolutionary biology, their application is still in its infancy in historical linguistics. This is in strong contrast to classical historical linguistics, where the quest for proto-forms and proto-meanings is often given more importance than the search for family trees and sub-groupings. In the following, we will report results of a pilot study on ancestral state reconstruction applied to lexicostatistical word list data. Our goal is to infer which words were used to *express* a given concept in the ancestral languages.

This task is not to be confused with *semantic reconstruction*, where linguists try to infer the original meaning of a given word. Instead, we try to infer the original word which expressed a given meaning. Since this approach reflects the onomasiological perspective on the linguistic sign, where the meaning of form-meaning pairs is fixed in order to allow to investigate how forms are used to conceptualize meanings over time, and since there does not exist a true term for this task, we chose the term "onomasiological reconstruction".[1] Classical semantic reconstruction in historical linguistics starts from a set of cognate words and tries to identify the original meaning of the ancestral word form (Wilkins, 1996). For this purpose, scholars try to take known directional tendencies into account. These tendencies are usually based

[1] We chose this term out of alternatives, not because we particularly like it, and we are aware that it may sound confusing for readers not too much acquainted with discussions on semantic change and lexical replacement, but we try to explain this in more detail below.

on the author's intuition, despite recent attempts to formalize and quantify the evidence (Urban, 2011). Following the classical distinction between *semasiology* and *onomasiology* in semantics, the former dealing with 'the meaning of individual linguistic expressions' (Bussmann, 1996, 1050), and the latter dealing with the question of how certain concepts are expressed (ibd. 834), semantic reconstruction is a *semasiological approach* to lexical change, as scholars start from the *meaning* of several lexemes in order to identify the meaning of the proto-form and its later development.

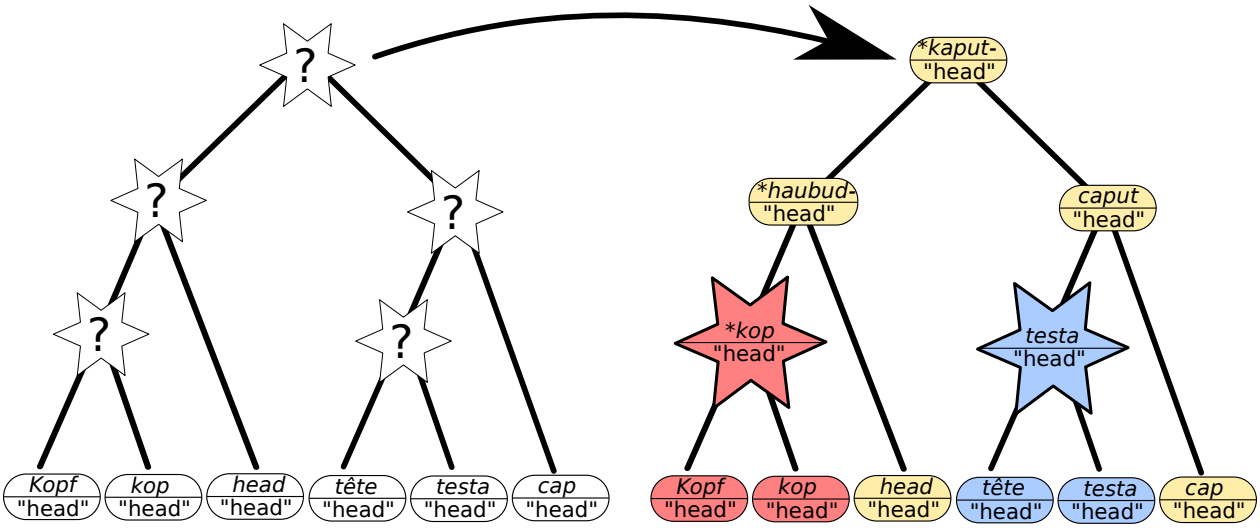

Figure 1: Ancestral state reconstruction: The graphic illustrates the key idea of ancestral state reconstruction. Given six words in genetically related languages, we want to know how these words evolved into their current shape. Having inferred a phylogeny of the languages as shown on the left of the figure, ancestral state reconstruction methods use this phylogeny to find the best way to explain how the six words have evolved along the tree, thereby proposing ancestral *states* of all words under investigation. The advantage of this procedure is that we can immediately identify not only the original nature of the characters we investigate, but also the changes they were subject to. Ancestral state reconstruction may thus yield important insights into historical processes, including sound change and lexical replacement

Instead of investigating lexical change from the semasiological perspective, one could also ask which of several word forms was used to denote a certain meaning in a given proto-language. This task is to some degree similar to proper semantic reconstruction, as it deals with the question of which meaning was attached to a given linguistic form. The approach, however, is *onomasiological*, as we start from the concept and search for the "name" that was attached to it. *Onomasiological semantic reconstruction*, the reconstruction of former *expressions*, has been largely ignored in classical semantic reconstruction.[2] This is unfortunate, since the onomasiological perspective may offer interesting insights into lexical change. Given that we are dealing with two perspectives on the same phenomenon, the onomasiological viewpoint may increase the evidence for semantic reconstruction.

[2]Notable exceptions include work of S. Starostin and collegues, compare, for example, Starostin (2016).

This is partially reflected in the "topological principle in semantic [i.e. onomasiological, GJ and JML] reconstruction" proposed by Kassian et al. (2015b). This principle uses phylogenies to support claims about the reconstruction of ancestral expressions in historical linguistics, trying to choose the 'most economic scenario' (ibd. 305) involving the least amount of semantic shifts. By adhering to the onomasiological perspective and modifying our basic data, we can model the problem of onomasiological reconstruction as an ancestral state reconstruction task, thereby providing a more formal treatment of the topological principle. In this task, we (1) start from a multilingual word lists in which a set of concepts has been translated into a set of languages (a classical "Swadesh list" or lexicostatistic word list, Swadesh 1955), (2) determine a plausible phylogeny for the languages under investigation, and (3) use ancestral state reconstruction methods to determine which word forms were most likely used to *express* the concepts in the ancestral languages in the tree. This approach yields an analysis as the one shown in Figure 1.

Although we think that such an analysis has many advantages over the manual application of the topological principle in onomasiological reconstruction employed by Kassian et al. (2015a), we should make very clear at this point that our reformulation of the problem as an ancestral state reconstruction task also bears certain shortcomings. First, since ancestral state reconstruction models character by character independently from each other, our approach relies on identical meanings only and cannot handle semantic fields with fine-grained meaning distinctions. This is a clear disadvantage compared to qualitative analyses, but given that models always simplify reality, and that neither algorithms nor datasets for testing and training are available for the extended task, we think it is justified to test how close the available ancestral state reconstruction methods come to human judgments. Second, our phylogenetic approach to onomasiological reconstruction does not answer any questions regarding semantic change, as we can only state which words are likely to have been used to express certain concepts in ancestral languages. This results clearly from the data and our phylogenetic approach, as mentioned before, and it is an obvious shortcoming of our approach. However, since the phylogenetic onomasiological reconstruction provides us with concrete hypotheses regarding the meaning of a given word on a given node in the tree, we can take these findings as starting point to further investigate how words changed their meaning afterwards. By providing a formal and data-driven way to apply the topological principle, we can certainly contribute to the broader tasks of semantic and onomasiological reconstruction in historical linguistics. As a third point, we should not forget that our method suffers the typical shortcomings of all data-driven disciplines, namely the shortcomings resulting from erroneous data assembly, especially erroneous cognate judgments, including undetected borrowings (Holm, 2007) and erroneous translations of the basic concepts which are investigated (Geisler and List, 2010) in all approaches based on lexicostatistical data. That errors in the data have an influence on the inferences made by the methods is obvious and clear. In order to make sure that we evaluate the full potential of phylogenetic

methods for ancestral state reconstruction, we therefore provide an exhaustive error analysis not only of the inferences made in our tests, but also of the data we used for testing.

In the following, we will illustrate how ancestral state reconstruction methods can be used to approximate onomasiological reconstruction in multilingual word lists. We test the methods on three publicly available datasets from three different language families and compare the results against experts' assessments.

2 Materials and Methods

2.1 Materials

2.1.1 Gold Standard

In order to test available methods for ancestral state reconstruction, we assembled lexical cognacy data from three publicly available sources, offering data on three different language families of varying size:

1. Indo-European languages, as reflected in the *Indo-European lexical cognacy database* (IELex, Dunn 2012, accessed from `http://ielex.mpi.nl/` on 9-5-2016),
2. Austronesian languages, as reflected in the *Austronesian Basic Vocabulary Database* (ABVD, Greenhill et al. 2008, accessed from `http://language.psy.auckland.ac.nz/austronesian/` on 12-2-2015), and
3. Chinese dialect varieties, as reflected in the *Basic Words of Chinese Dialects* (BCD, Wang 2004).

All datasets are originally classical word lists as they are used in standard approaches to phylogenetic reconstruction: They contain a certain number of concepts which are translated into the target languages and are then annotated for cognacy. In order to be applicable as a test set for our analysis, the datasets further need to list proto-forms of the supposed ancestral language of all languages in the sample. All data that we used for our studies is available from the supplementary material.

The **BCD** database was underlying the study of Ben Hamed and Wang (2006) and is no longer accessible via its original URL, but it has been included as part of the publication by List (2015) and later been revised in List (2016). It comprises data on 200 basic concepts (a modified form of the concept list by Swadesh 1952) translated into 23 Chinese dialect varieties. Additionally Wang (2004) lists 230 translations in Old Chinese for 197 of the 200 concepts. Since Old Chinese is the supposed ancestor of all Chinese dialects, this data qualifies as a gold standard for our experiment on ancestral state reconstruction. We should, however, bear in mind that the relationship between Old Chinese, as a variety spoken some time between 800 and 200 BC and the most recent common ancestor of all Chinese dialects, spoken

between 200 and 400 CE, is a remote one. We will discuss this problem in more detail in our linguistic evaluation of the results in section 4. Given that many languages contain multiple synonyms for the same concept, the data, including Old Chinese, comprises 5,437 words, which can be clustered into 1,576 classes of cognate words, of which 980 are "singletons", that is, they comprise classes containing only one single element. Due to the large time span between Old Chinese and the most recent common ancestor of all Chinese dialects, not all Old Chinese forms are technically reconstructible from the data, as they reflect words that have been lost in all dialects. As a result, we were left with 144 reconstructible concepts for which at least one dialect retains an ancestral form attested in Old Chinese.

For the **IELex** data,[3] we used all languages and dialects except those marked as "Legacy" and two creole languages (*Sranan* and *French Creole Dominica*, as lexical change arguably underlies different patterns than in normal language change under creolization). This left us with 134 languages and dialects, including 31 ancient languages (*Ancient Greek, Avestan, Classical Armenian, Gaulish, Gothic, Hittite, Latin, Luvian, Lycian, Middle Breton, Middle Cornish, Mycenaean Greek, Old Persian, Old Prussian, Old Church Slavonic, Old Gutnish, Old Norse, Old Swedish, Old High German, Old English, Old Irish, Old Welsh, Old Cornish, Old Breton, Oscan, Palaic, Pali, Tocharian A, Tocharian B, Umbrian, Vedic Sanskrit*). The data contain translations of 208 concepts into those languages and dialects (often including several synonymous expressions for the same concept from the same language). Most entries are assigned a *cognate class label*. We only used entries containing an unambiguous class label, which left us with 26,524 entries from 4,352 cognate classes. IELex also contains 167 reconstructed entries (for 135 concepts) for Proto-Indo-European. These reconstructions were used as Gold Standard to evaluate the automatically inferred reconstructions.

ABVD contains data from a total of 697 Austronesian languages and dialects. We selected a subset of 349 languages (all taken from the 400-language sample used in Gray et al. 2009), each having a different ISO code which is also covered in the Glottolog database (Hammarström et al., 2015). ABVD covers 210 concepts, with a total of 44,983 entries from 7,727 cognate classes for our 349-language sample. It also contains 170 reconstructions for Proto-Austronesian (each denoting a different concept) including cognate-class assignments. An overview of the data used is given in Table 1.

Dataset	Languages	Concepts	Cognate Classes	Singletons	Words
IELex	134	207 (135 reconstructible)	4,352	1,434 singletons	26,524
ABVD	349	210 (170 reconstructible)	7,727	2,671 singletons	44,983
BCD	24	200 (144 reconstructible)	1.576	980 singletons	5,437

Table 1: Datasets used for ancestral state reconstruction. "Reconstructible" states in the column showing the number of concepts refer to the amount of concepts in which the proto-form is reflected in at least one of the descendant languages. "Singletons" refer to cognate sets with only one reflex, which are not informative for the purpose of certain methods of ancestral state reconstruction, like the MLN approach, and therefore excluded from the analysis.

[3] IELex is currently being thoroughly revised as part of the *Cognates in the Basic Lexicon* (COBL) project, but since this data has not yet been publicly released, we were forced to use the IELex data which we retrieved from `ielex.mpi.nl`.

2.2 Methods

2.2.1 Reference Phylogenies

All ASR methods in our test (except the baseline) rely on phylogenetic information when inferring ancestral states, albeit to a different degree. Some methods operate on a single tree topology only, while other methods also use branch lengths information or require a sample of trees to take phylogenetic uncertainty into account. To infer those trees, we arranged the cognacy information for each data set into a presence-absence matrix. Such a data structure is a table with languages as rows and cognate classes occurring within the data set as columns. A cell for language *l* and cognate class *cc* for concept *c* has entry

- 1 if *cc* occurs among the expressions for *c* in *l*,
- 0 if the data contain expressions for *c* in *l*, but none of them belongs to *cc*, and
- undefined if *l* does not contain any expressions for *c*.

Bayesian phylogenetic inference was performed on these matrices. For each data set, tree search was constrained by *prior* information derived from the findings of traditional historical linguistics. More specifically, we used the following prior information:

- **IELex.** We used 14 topological constraints (see Figure 2), age constraints for the 31 ancient languages, and age constraints for 11 of the 14 topological constraints.

  The age constraints for *Middle Breton, Middle Cornish, Mycenaean Greek, Old Breton, Old Cornish, Old Welsh*, and *Palaic* are based on information from `http://multitree.org/` (accessed on 10-14-2016). The age constraint for *Pali* is based on information from `http://www.britannica.com` (accessed on 10-14-2016). The constraints for *Old Gutnish* are taken from (Wessen, 1968) and those for *Old Swedish* and *Old High German* from (Campbell and King, 2013). All other age constraints are derived from the Supplementary Information of (Bouckaert et al., 2012).

- **ABVD.** We only considered trees consistent with the Glottolog expert classification (Hammarström et al., 2015). This amounts to 213 topological constraints.

- **BDC.** We only considered trees consistent with the expert classification from Sagart (2011). This amounts to 20 topological constraints.

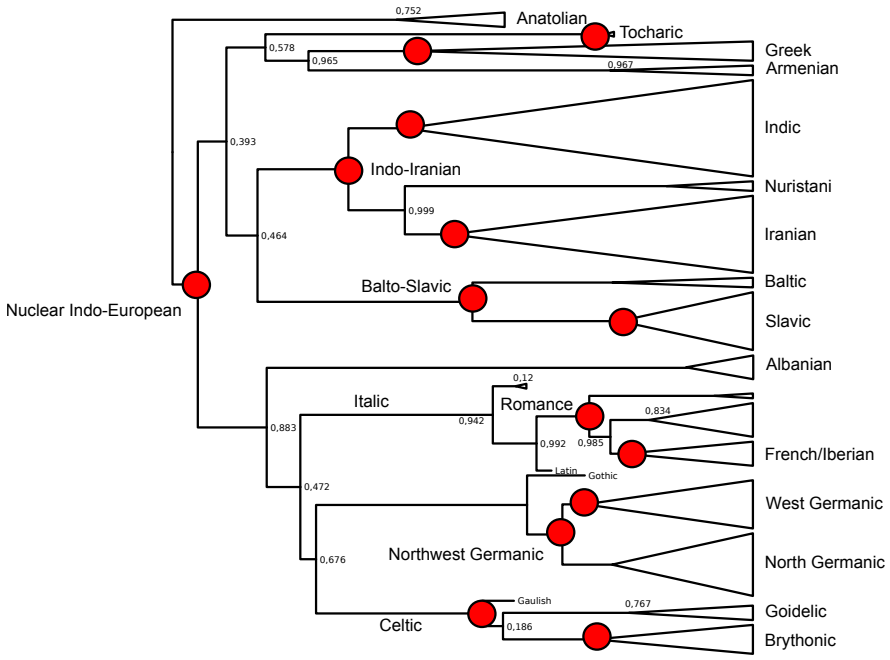

Figure 2: Maximum Clade Credibility tree for IELex (schematic). Topological constraints are indicated by red circles. Numbers at intermediate nodes indicate posterior probabilities (only shown if < 1).

Analyses were carried out using the MrBayes software (Ronquist et al., 2012). Likelihoods were computed using ascertainment bias correction for all-absent characters and assuming Gamma-distributed rates (with 4 Gamma categories). Regarding the tree prior, we assumed a relaxed molecular clock model (more specifically, the *Independent Gamma Rates* model, cf. Lepage et al. 2007, with an exponential distribution with rate 200 as prior distribution for the variance of rate variation). Furthermore we assumed a birth-death model (Yang and Rannala, 1997) and random sampling of taxa with a sampling probability of 0.2. For all other parameters of the prior distribution, the defaults offered by the software were used.[4]

For each dataset, a *maximum clade credibility tree* was identified as the **reference tree** (using the software *TreeAnnotator*, `http://beast2.org/treeannotator/`, retrieved on September 13, 2016; part of the software suite Beast, cf. Bouckaert et al. 2014). Additionally, 100 trees were sampled from the posterior distribution for each dataset and used as **tree sample** for ASR.

[4]These defaults are: uniform distribution over equilibrium state frequencies; standard exponential distribution as prior for the shape parameter α of the Gamma distribution modeling rate variation; standard exponential distribution as prior over the tree age, measured in expected number of mutatations per character.

2.2.2 Ancestral State Reconstruction

For our study, we tested three different established **algorithms**, namely (1) Maximum Parsimony (MP) reconstruction using the Sankoff algorithm (Sankoff, 1975), (2) the minimal lateral network (MLN) approach (Dagan et al., 2008) as a variant of Maximum Parsimony in which parsimony weights are selected with help of the *vocabulary size criterion* (List et al., 2014b,c), and (3) Maximum Likelihood (ML) reconstruction as implemented in the software *BayesTraits* (Pagel and Meade, 2014). These algorithms are described in detail below.

We tested two different ways to arrange cognacy information as **character matrices**:

- **Multistate characters.** Each concept is treated as a character. The value of a character for a given language is the cognate class label of that language's expression for the corresponding concept. If the data contain several non-cognate synonymous expressions, the language is treated as polymorphic for that character. If the data do not contain an expression for a given concept and a given language, the corresponding character value is undefined.
- **Binary characters.** Each cognate class label that occurs among the documented languages of a dataset is a character. Possible values are 1 (a language contains an expression from that cognate class), 0 (a language does not contain an exponent of that cognate class, but other expressions for the corresponding concept are documented) or *undefined* (the data do not contain an expression for the concept in question from the language in question).

All three algorithms rely on a reference phylogeny to infer ancestral states. To test the impact of **phylogenetic uncertainty**, we performed ASR both on the *reference tree* and on the *tree sample* for all three algorithms.

Maximum Parsimony (MP) A *complete scenario* for a character is a phylogenetic tree where all nodes are labeled with some character value. For illustration, three scenarios are shown in Figure 3. The *parsimony score* of a scenario is the number of mutations, i.e., of branches where the mother node and the daughter node carry different labels. Now suppose only the labels at the tips of the tree are given. The parsimony score of such a *partial scenario* is the minimal parsimony score of any complete scenario consistent with the given tip labels. In the example in Figure 3, this value would be 2. The ASR for the root of the tree would be the root label of the complete scenario giving rise to this minimal parsimony score. If several complete scenarios with different root labels give rise to the same minimal score, all their root labels are possible ASRs. This logic can be generalized to *weighted parsimony*. In this framework, each mutation from a state at the mother node to the state at the daughter node of a tree has a certain *penalty*, and these penalties may differ for different types of mutations. The overall parsimony score of a complete scenario is sum of all penalties for all mutations in this scenario.[5]

[5]There is a variant of MP called *Dollo parsimony* (Le Quesne, 1974; Farris, 1977) which is *prima facie* well-suited for modeling cognate class evolution. Dollo parsimony rests on the assumption that complex characters evolve only once, while they may be lost multiple times. If "1" represents

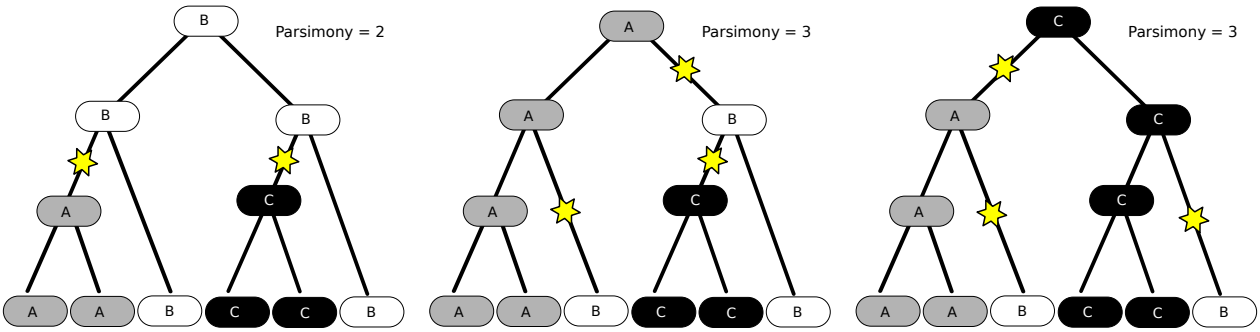

Figure 3: Complete character scenarios. Mutations are indicated by yellow stars.

The *Sankoff algorithm* is an efficient method to compute the parsimony score and the root ASR for a partial scenario. It works as follows. Let *states* be the ordered set of possible states of the character in question, and let n be the cardinality of this set. For each pair of states i, j, $w(i, j)$ is the penalty for a mutation from states_i to states_j.

- **Initialization.** Each tip t of the tree is initialized with a vector $\mathrm{wp}(t)$ of length n, with $\mathrm{wp}(t)_i = 0$ if t's label is states_i, and ∞ else. (If t is polymorphic, all labels occuring at t have the score 0.)

- **Recursion.** Loop through the non-tip nodes of the tree bottom-up, i.e., visit all daughter nodes before you visit the mother node. Each non-terminal node *mother* with the set *daughters* as daughter nodes is annotated with a vector $\mathrm{wp}(\mathit{mother})$ according to the rule

$$\mathrm{wp}(\mathit{mother})_i \quad = \quad \sum_{d \in \mathit{daughters}} \min_{1 \leq j \leq n} \left(w(i,j) + \mathrm{wp}(d)_j \right)$$

- **Termination.** The parsimony score is $\min_{1 \leq i \leq n} \mathrm{wp}(\mathit{root})_i$ and the root ASR is $\arg\min_{1 \leq i \leq n} \mathrm{wp}(\mathit{root})_i$.

If MP-ASR is performed on a sample of trees, the Sankoff algorithm is applied to each tree in the sample, and the vectors at the roots are summed up. The root-ASR is then the state with the minimal total score. For our experiments, we used the following **weight matrices**:

presence and "0" absence of such a complex character, the weight of a mutation $1 \rightarrow 0$ should be infinitesimally small in comparison to the weight of $0 \rightarrow 1$. Performing ASR under this assumption amounts to projecting each character back to the latest common ancestor of all its documented occurrences. While this seems initially plausible since each cognate class can, by definition, emerged only once, recent empirical studies have uncovered that multiple mutations $0 \rightarrow 1$ can easily occur with cognate-class characters. A typical scenario are parallel semantic shifts. Chang et al. (2015), e.g., point out that descendent words of Proto-Indo-European **pod-* 'foot' independently shifted their meaning to 'leg' both in Modern Greek and in Modern Indic and Iranian languages. So the Modern Greek *πόδι* and the Marathi *pāy*, both meaning 'leg', are, according to IELex, cognate, but the latest common ancestor language of Greek and Marathi (Nuclear Proto-Indo-European or a close descendant of it) probably used a non-cognate word to express 'leg'. Other scenarios leading to the parallel emergence of cognate classes are loans and *incomplete lineage sorting*; see the discussion in Section 4. Bouckaert et al. (2012) test a probabilistic version of the Dollo approach and conclude that a time-reversible model provides a better fit of cognate-class character data.

- For multistate characters, we used uniform weights, i.e., $w(i,i) = 0$ and $w(i,j) = 1$ iff $i \neq j$.
- For binary presence-absence characters, we assumed that the penalty of a gain is twice as high as the penalty for a loss: $w(i,i) = 0$, $w(1,0) = 1$, and $w(0,1) = 2$.[6]

For a given tree and a given character, the Sankoff algorithm produces a parsimony score for each character state. If the cognacy data are organized as multi-state characters, each state is a cognate class. The *reconstructed states* are those achieving the minimal value among these scores. If a tree sample, rather than a single tree, is considered, the parsimony scores are averaged over the resuls for all trees in the sample. The reconstructed states are those achieving the minimal average score. If the cognacy data are organized as presence-absence characters, we consider the parsimony scores of state "1" for all cognate classes expressing a certain concept. The reconstructed cognate classes are those achieving the minimal score for state "1". If a tree sample is considered, scores are averaged over trees.

Minimal Lateral Networks (MLN) The MLN approach was originally developed for the detection of lateral gene transfer events in evolutionary biology (Dagan et al., 2008). In this form, it was also applied to linguistic data (Nelson-Sathi et al., 2011), and later substantially modified (List et al., 2014b,c). While the original approach was based on very simple gain-loss-mapping techniques, the improved version uses weighted parsimony on presence-absence data of cognate set distributions. In each analysis, several parameters (ratio of weights for gains and losses) are tested, and the best method is then selected, using the criterion of *vocabulary size distributions*, which essentially claims that the amount of synonyms per concept in the descendant languages should not differ much from the amount of synonyms reconstructed for ancestral languages. Thus, of several competing scenarios for the development of characters along the reference phylogeny, the scenario that comes closest to the distribution of words in the descendant languages is selected. This is illustrated in Figure 4. Note that this criterion may make sense intuitively, if one considers that a language with excessive synonymy would make it more difficult for the speakers to communicate. Empirically, however, no accounts on average synonym frequencies across languages are available, and as a result, this assumption remains to be proven in future studies.

While the improved versions were also essentially used to infer borrowing events in linguistic datasets, List (2015) showed that the MLN approach can also be used for the purpose of ancestral state reconstruction, given that it is based on a variant of weighted parsimony. Describing the method in all its details would go beyond the scope of this paper. For this reason, we refer the reader to the original publications which describe the algorithm in detail, as well as the actual source code which is published along with the LingPy software package (`http://lingpy.org`, List and Forkel

[6]The ratio between gains and losses follows from the experience with the MLN approach which is presented in more detail below and which essentially tests different gain-loss scenarios for their suitability to explain a given dataset. In all published studies in which the MLN approach was tested (List et al., 2014b,c; List, 2015), the best gain-loss ratio reported was 2:1.

2016). To contrast MLN against the variant of Sankoff parsimony we used, it is, however, important to note that the MLN method does not handle *singletons* in the data, that is, words which are not cognate with any other words.[7] It should also be kept in mind that the MLN method in its currently available implementation only allows for the use of *binary characters states*: multi-state characters are not supported and can therefore not be included in our test.

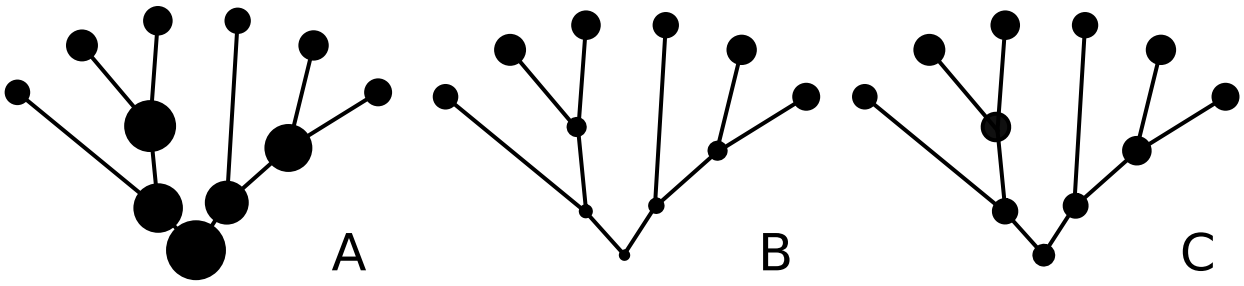

Figure 4: Vocabulary Size Distributions as a criterion for parameter selection in the MLN approach. A shows an analysis which proposes far too many words in the ancestral languages, B proposes far to few words, and C reflects and optimal scenario.

Maximum Likelihood (ML) While the Maximum Parsimony principle is conceptually simple and appealing, it has several shortcomings. As it only uses topological information and disregards branch lengths, it equally penalizes mutations on short and on long branches. However, mutations on long branches are intuitively more likely than on short branches if we assume that branch lengths corresponds to historical time. Also, MP entirely disregards the possibility of multiple mutations on a single branch. It would go beyond the scope of this article to fully spell out the ML method in detail; the interested reader is referred to the standard literature on phylogenetic inference (such as Ewens and Grant 2005, Section 15.7) for details. In the sequel we will confine ourselves to presenting the basic ideas.

The fundamental assumption underlying ML is that character evolution is a *Markov process.* This means that mutations are non-deterministic, stochastic events, and their probability of occurrence only depend on the current state of the language. For simplicity's sake, let us consider only the case where there are two possible character states, 1 (for presence of a trait) and 0 (absence). Then there is a probability p_{01} that a language gains the trait within one unit of time, and p_{10} that it looses it.

[7]The technical question of parsimony implementations is here, whether one should penalize the origin of a character in the root or not. The parsimony employed by MLN penalizes all origins. As a result, words which are not cognate with any other word can never be reconstructed to a node higher in the tree. For a discussion of the advantages and disadvantages of this treatment, see Mirkin et al. (2003).

The probability that a language switches from state i to state j within a time interval t is then given by the *transition probability* $P(t)_{ij}$:[8]

$$\begin{aligned} \alpha &= \frac{p_{01}}{p_{01}+p_{10}} \\ \beta &= \frac{p_{10}}{p_{01}+p_{10}} \\ \lambda &= -\log(1-p_{01}-p_{10}) \\ P(t) &= \begin{pmatrix} \beta+\alpha e^{-\lambda t} & \alpha-\alpha e^{-\lambda t} \\ \beta-\beta e^{-\lambda t} & \alpha+\beta e^{-\lambda t} \end{pmatrix} \end{aligned}$$

α and β are the *equilibrium probabilities* of states 1 and 0 respectively, and λ is the *mutation rate*. If t is large in comparison to the minimal time step (such as the time span of a single generation), we can consider t to be a continuous variable and the entire process a *continuous time Markov process*. The behavior of this process is illustrated in Figure 5 for $\alpha = 0.2$, $\beta = 0.8$, and $\lambda = 1$. If a language is in state 0 at time 0, its probability to be in state 1 after time t is

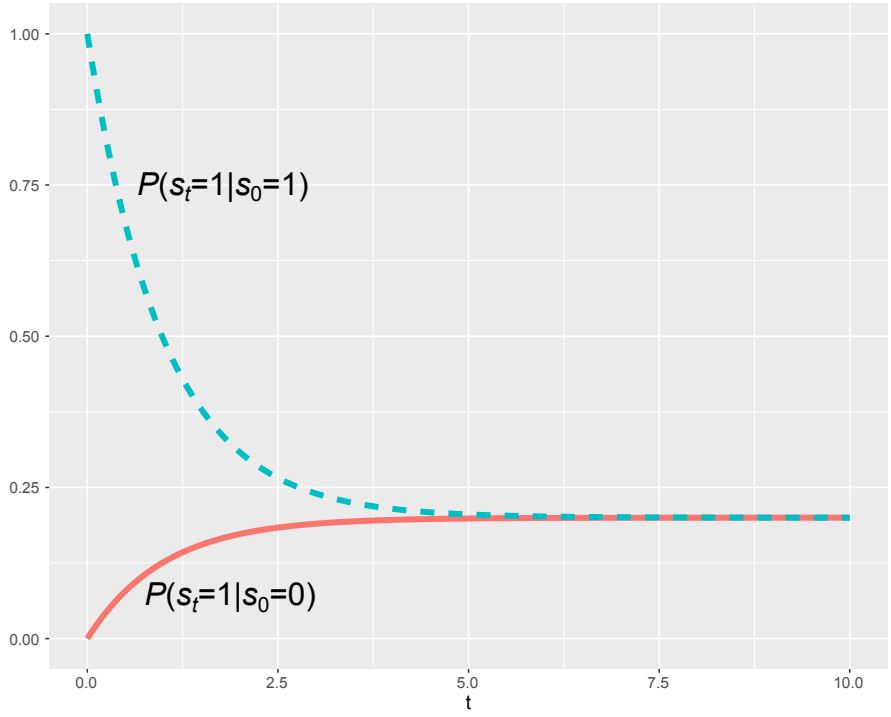

Figure 5: Gain and loss probabilities under a continuous-time Markov process

[8] We assume that the rows and columns of $P(t)$ are indexed with $0, 1$.

indicated by the solid line. This probability continuously increases and converges to α. This is the gross probability to start in state 0 and end in state 1; it includes the possibility of multiple mutations, as long as the number of mutations is odd. The dotted line shows the probability of ending up in state 1 after time t when a language starts in state 1. This quantity is initally close to 100%, but it also converges towards α over time. In other words, the absence of mutations (or a sequence of mutations that re-established the initial state) is predicted to be unlikely over long periods of time. In a complete scenario, i.e., a phylogenetic tree with labeled non-terminal nodes, the likelihood of a branch is the probability of ending in the state at the daughter node if one starts in the state of the mother node after a time interval given by the branch length.

The overall likelihood of a complete scenario is the product of all branch likelihoods, multiplied with the equilibrium probability of its root state. The likelihood of a partial scenario, where only the states of the tips are known, is the sum of the likelihoods of all complete scenarios consistent with it. It can efficiently be computed in a way akin to the Sankoff algorithm. ($\mathcal{L}(x)$ is the likelihood vector of node x, and π_i is the equilibrium probability of state i).

- **Initialization.** Each tip t of the tree is initialized with a vector $\mathcal{L}(t)$ of length n, with $\mathcal{L}(t)_i = 1$ if t's label is $states_i$, and 0 else. (If t is polymorphic, all labels occuring at t have the same likelihood, and these likelihoods sum up to 1.)

- **Recursion.** Loop through the non-tip nodes of the tree bottom-up, i.e., visit all daughter nodes before you visit the mother node. Each non-terminal node *mother* with the set *daughters* as daughter nodes is annotated with a vector $\mathcal{L}(mother)$ according to the rule

$$\mathcal{L}(mother)_i \quad = \quad \prod_{d \in daughters} \sum_{1 \leq j \leq n} (P(t)(i,j)\mathcal{L}(d)_j),$$

  where t is the length of the branch connecting d to its mother node.

- **Termination.** The likelihood of the scenario is $\sum_{1 \leq i \leq n} \mathcal{L}(root)_i$. The ASR likelihood of state i is proportional to $\pi_i \mathcal{L}(root)_i$.[9]

The likelihood of the scenario calculated this way is the sum of the likelihoods of all scenarios compatible with the information at the tips. The overall likelihood of a tree for a character matrix is the product of the likelihoods for the individual characters. (This captures the simplifying assumption that characters are mutually stochastically independent.)

[9]Note that this approach can only be used to compute the *marginal likelihood* of states at the *root of the tree.* To perform ASR at interior nodes or joint ASR at several nodes simultaneously, a more complex approach is needed. These issues go beyond the scope of this article though.

As the model parameters (λ and the equilibrium probabilities) are not known *a priori*, they are estimated from the data. This is done by choosing values that maximize the overall likelihood of the tree for the given character matrix, within certain constraints. In our experiments we used the following constraints:

- For multistate characters, we assumed a uniform equilibrium distribution for all characters, and identical rates for all character transitions.
- For binary characters, we assume equilibrium probabilities to be identical for all characters. Those equilibrium probabilities were estimated from the data as the empirical frequencies. We assumed *gamma-distributed rates*, i.e., rates were allowed to vary to a certain degree between characters.

Once the model parameters are fixed, the algorithm produces a probability distribution over possible states for each character. The *reconstructed states* are identified in a similar way as for Sankoff parsimony. First these probabilities are averaged over all trees if more than one tree is considered. For multistate characters, the state(s) achieving the highest probability are selected. For binary presence-absence characters, those cognate classes for a given concept are selected that achieve the highest average probability for state 1.

2.3 Evaluation

For all three data sets considered, the Gold Standard contains cognate class assignments for a common ancestor language. For the Chinese data, these are documented data for Old Chinese. For the other two datasets, these are reconstructed forms of the supposed latest common ancestor (LCA), Proto-Indo-European and Proto-Austronesian respectively. The Old Chinese variety is not identical with the latest common ancestor of all Chinese dialects but rather predates it by several hundred years. Due to the rather stable character of the written languages as opposed to the vernaculars throughout the history of Chinese, it is difficult to assess with certainty which exact words were used to denote certain basic concepts, and Old Chinese as reflected in classical sources is a compromise solution as it allows us to consider written evidence rather than reconstructed forms (see section 4 for a more detailed discussion).

For the evaluation, we only consider those concepts for which (a) the LCA data identify a cognate class and (b) this cognate class is also present in one of the descendant languages considered in the experiment. The Gold Standard defines a set of cognate classes that were present in the LCA language. Let us call this set *LCA*. Each ASR algorithm considered defines a set of cognate classes that are reconstructed for the LCA. We denote this set as *ASR*. In the sequel we will deploy evaluation metrics established in machine learning to assess how well these two sets coincide:

$$
\begin{aligned}
\textit{precision} &\doteq \frac{|LCA \cap ASR|}{|ASR|} \\
\textit{recall} &\doteq \frac{|LCA \cap ASR|}{|LCA|} \\
\textit{F-score} &\doteq 2 \times \frac{\textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}
\end{aligned}
$$

The *precision* expresses the proportion of correct reconstructions among all reconstructions. The *recall* gives the proportion of ancestral cognate classes that are correctly reconstructed. The *F-score* is the harmonic mean between precision and recall.

Results for the various ASR algorithms are compared against a **frequency baseline**. According to the baseline, a cognate class cc for a given concept c is reconstructed if and only if cc occurs at least as frequently among the languages considered (excluding the LCA language) as any other cognate class for c. This baseline comes very close to the current practice in classical historical linguistics, as presented in Starostin (2016), although it is clear that trained linguists practicing onomasiological reconstruction may take many additional factors into account. For IELex, we also considered a second baseline, dubbed the **sub-family baseline**. A cognate class cc is deemed reconstructed if and only if it occurs in at least two different sub-families, where sub-families are *Albanian, Anatolian, Armenian, Balto-Slavic, Celtic, Germanic, Greek, Indo-Iranian, Italic*, and *Tocharian*.

3 Results

The individual results for all data sets and algorithm variants are given in Tables 2, 3 and 4. Note that MLN does not offer a multi-state variant, so for MLN, only results for binary states are reported. The effects of the various design choices — coding characters as multi-state of binary; using a single reference tree or a sample of trees — as well as the differences between the three ASR algorithms considered here are summarized in Figure 6. The bars represent the average difference in F-score to the frequency baseline, averaged over all instances of the corresponding category across data sets.

It is evident that there are major differences in the performance of the three algorithms considered. While the F-score for MLN-ASR remains, on average, below the baseline, Sankoff-ASR and ML-ASR clearly outperform the baseline. Furthermore, ML-ASR clearly outperforms Sankoff-ASR. Given that both MLN-ASR and Sankoff-ASR deal with Maximum Parsimony, the rather poor performance of the MLN approach shows that the basic vocabulary size criterion may not be the best criterion for penalty selection in parsimony approaches. It may also be related to further individual

algorithm	characters	tree	precision	recall	F-score
frequency baseline	*multi*	-	0.599	0.590	0.594
MLN	*bin*	*single*	0.568	0.729	0.638
MLN	*bin*	*sample*	0.568	0.729	0.638
Sankoff	*multi*	*single*	0.484	0.743	0.586
Sankoff	*multi*	*sample*	0.510	0.722	0.598
Sankoff	*bin*	*single*	0.596	0.688	0.639
Sankoff	*bin*	*sample*	0.651	0.660	0.655
ML	*multi*	*single*	0.669	0.660	**0.664**
ML	*multi*	*sample*	0.669	0.660	**0.664**
ML	*bin*	*single*	0.634	0.625	0.629
ML	*bin*	*sample*	0.641	0.632	0.636

Table 2: Evaluation results for Chinese

algorithm	characters	tree	precision	recall	F-score
frequency baseline	*multi*	-	0.607	0.497	0.547
sub-family baseline	*bin*	-	0.402	0.885	0.553
MLN	*bin*	*single*	0.781	0.303	0.437
MLN	*bin*	*sample*	0.781	0.303	0.437
Sankoff	*multi*	*single*	0.367	0.739	0.491
Sankoff	*multi*	*sample*	0.566	0.594	0.580
Sankoff	*bin*	*single*	0.542	0.630	0.583
Sankoff	*bin*	*sample*	0.597	0.503	0.546
ML	*multi*	*single*	0.741	0.606	0.667
ML	*multi*	*sample*	0.763	0.624	0.687
ML	*bin*	*single*	0.778	0.636	0.700
ML	*bin*	*sample*	0.785	0.642	**0.707**

Table 3: Evaluation results for IELex

choices introduced in the MLN algorithm or our version of Sankoff parsimony. Given that the MLN approach was not primarily created for the purpose of ancestral state reconstruction, our findings do not necessarily invalidate the approach, yet they show that it might be worthwhile to further improve on its method for ancestral state reconstruction.

The impact of the other choices is less pronounced. Binary character coding provides on average slightly better results than multistate character coding, but the effect is minor. Likewise, capturing information about phylogenetic uncertainty by using a sample of trees leads, on average, to a slight increase in F-scores, but this effect is rather small as well.

To understand why ML is superior to the two parsimony-based algorithms tested here, it is important to consider the conceptual differences between parsimony-based and likelihood-based ASR. Parsimony-based approaches operate on the tree topology only, disregarding branch lengths. Furthermore the numerical parameters being used, i.e. the mutation

algorithm	characters	tree	precision	recall	F-score
frequency baseline	*multi*	-	0.618	0.618	0.618
MLN	*bin*	*single*	0.843	0.412	0.553
MLN	*bin*	*sample*	0.882	0.394	0.545
Sankoff	*multi*	*single*	0.688	0.849	0.760
Sankoff	*multi*	*sample*	0.726	0.816	0.768
Sankoff	*bin*	*single*	0.723	0.771	0.746
Sankoff	*bin*	*sample*	0.757	0.749	0.753
ML	*multi*	*single*	0.788	0.788	**0.788**
ML	*multi*	*sample*	0.788	0.788	**0.788**
ML	*bin*	*single*	0.776	0.776	0.776
ML	*bin*	*sample*	0.771	0.771	0.771

Table 4: Evaluation results for ABVD

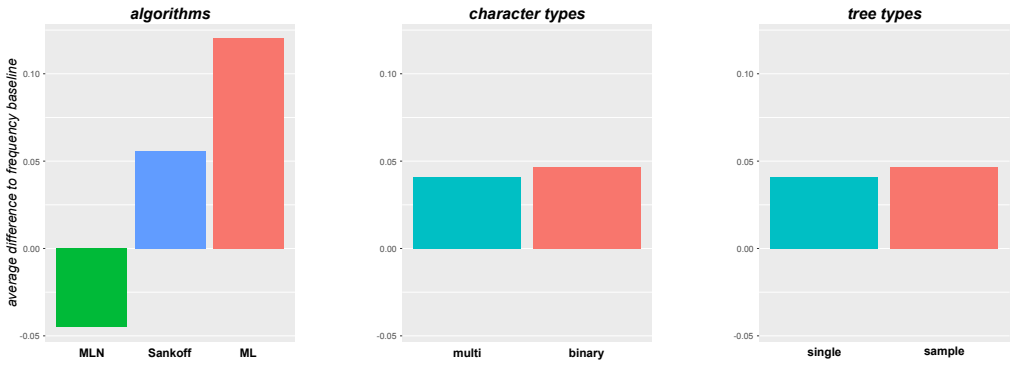

Figure 6: Average differences in F-score to frequency baseline

penalties, are fixed by the researcher based on intuition and heuristics. ML, in contradistinction, uses branch length information, and it is based on an explicit probabilistic model of character evolution.

This point is illustrated in Figure 7, which schematically displays ASR for the concept *eat* for the Chinese dialect data. The left panel visualizes Sankoff ASR and the right panel Maximum-Likelihood ASR. The guide tree identifies two sub-clades, shown as the upper and lower daughter of the root node. The dialects in the upper part of the tree represent the large group of Northern and Central dialects, including the dialect of Beijing, which comes close to standard Mandarin Chinese. The dialects in the lower part of the tree represent the diverse Southern group including the archaic Mǐn 闽 dialects spoken at the South-Eastern coast, as well as well as Hakka, and Yuè 粵, sometimes also called Cantonese, the prevalent variety spoken in Hong Kong. All Southern dialects use the same cognate class (*eat.Shi.1327*, Mandarin Chinese *shí* 食, nowadays only reflected in compounds) and all Northern and Central dialects use a different cognate class (*eat.Chi.243*, Mandarin Chinese *chī* 吃, regular word for 'eat' in most Northern varieties). Not surpris-

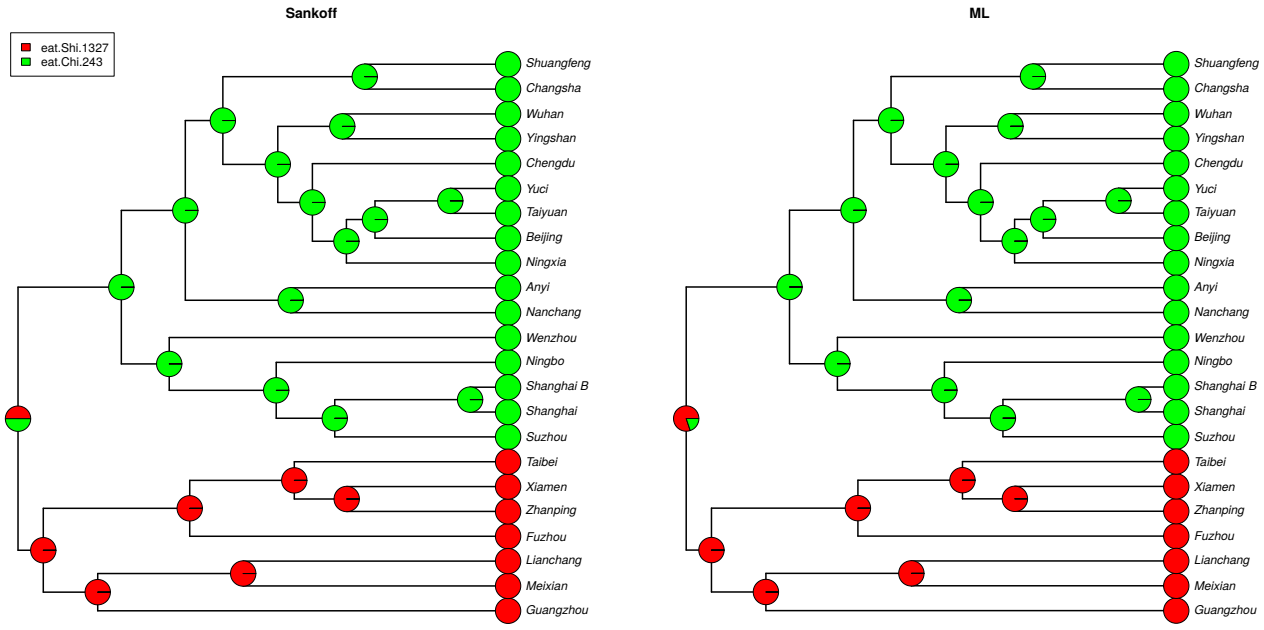

Figure 7: Maximum-Likelihood ASR and Sankoff Parsimony ASR for the concept *eat* for Chinese dialect data

ingly, both algorithms reconstruct *eat.Shi.1327* to the ancestor of the Southern dialects and *eat.Chi.243* to ancestor of the Northern and Central dialects. Sankoff ASR only uses the tree topology to reconstruct the root state. Since the situation is entirely symmetric regarding the two daughters of the root, the two cognate classes are tied with exactly the same parsimony score at the root. Maximum-Likelihood ASR additionally takes branch lengths into account. Since the latest common ancestor of the Southern dialects is closer to the root than the latest common ancestor of the Northern and Central dialects, the likelihood of a mutation along the lower branch descending from the root is smaller than along the upper branch. Therefore the lower branch has more weight when assigning probabilities to the root state. Consequently, *eat.Shi.1327* comes out as the most likely state at the root – which is in accordance with the gold standard. Our findings indicate that the more fine-grained, parameter rich Maximum-Likelihood approach is generally superior to the simpler parsimony-based approaches.

The parameters of the Maximum-Likelihood model, as well as the branch lengths, are estimated from the data. Our findings underscore the advantages of an empiricist, stochastic and data-driven approach to quantitative historical linguistics as compared to more heuristic and parameter-poor methods.

4 Linguistic Evaluation of the Results

The evaluation of the results against a gold standard can help us to understand the general performance of a given algorithm. Only a careful linguistic evaluation, however, helps us to understand the concrete difficulties and obstacles that the algorithms have to face when being used to analyze linguistic data. In order to account for this we carried out detailed linguistic evaluations of the results proposed for **IELex** and **BCD**. In these evaluations, we compared the results of the best methods (Binary ML Sample for IELex, and Multi ML for BCD) for each of the datasets with the respective gold standards, searching for potential reasons for the differences between automatic method and gold standard. In each of the two evaluations which are given in Appendix B.1 and B.2, we compared those forms which were reconstructed back to the root in the gold standard but missed by the gold standard, and those forms proposed by the algorithm but not by the gold standard. By consulting additional literature and databases, we could first determine whether the error was due to the algorithm or due to a problem in the gold standard. In a next step, we tried to identify the most common sources of errors, which we assigned to different error classes. Due to the differences in the histories and the time depths, the error-classes we identified slightly differ, and while a rather common error in IELex were erroneous cognate judgments in the gold standard,[10] we find many problematic meanings in BCD which are rarely expressed overtly in Chinese dialects.[11] Apart from errors which were hard to classify and thus not assigned to any error-class, problems resulting from the misinterpretation of branch-specific cognate sets, as well as problems resulting from parallel semantic shift (homoplasy) were among the most frequently occurring problems in both datasets.

Figure 8 gives detailed charts of the error analyses for missed and erroneously proposed items in the two datasets. The data is listed in such a way that mismatches between gold standard and algorithms can be distinguished. When inspecting the findings for IELex, we can thus see that the majority of the 59 missed cognates can be attributed to cognate sets which are only reflected in one branch in the Indo-European languages and do therefore not qualify as good candidates to be reconstructed back to the proto-language. As an example, consider the form *pneu̯- (cognate class `breathe:P`) which is listed as onomasiological reconstruction for the concept ‘to breathe’ in the gold standard. As it only occurs in Ancient Greek and has no reflexes in any other language family, this root is highly problematic, as is also confirmed by the *Lexicon of Indo-European Verbs*’ where the root is flagged as questionable (Rix et al., 2001, 489). Second, in the error-statistics for Indo-European are cognate sets whose onomasiological reconstruction is not confirmed by plausible semantic reconstructions in the gold standard. As an example for this error class, consider the form *dhōĝh-e/os- (cognate class `day:B`) proposed for the meaning slot ‘day’. While Kroonen (2013, 86f) confirms the reconstruction of the root, as it occurs in Proto-Germanic and Indo-Iranian, the meaning ‘day’ is by no means clear,

[10] See Appendix B.1 for details

[11] Examples include meanings for ‘if’, ‘because’, etc., which may be expressed but may as well be omitted in normal speech, see Appendix B.2 for details.

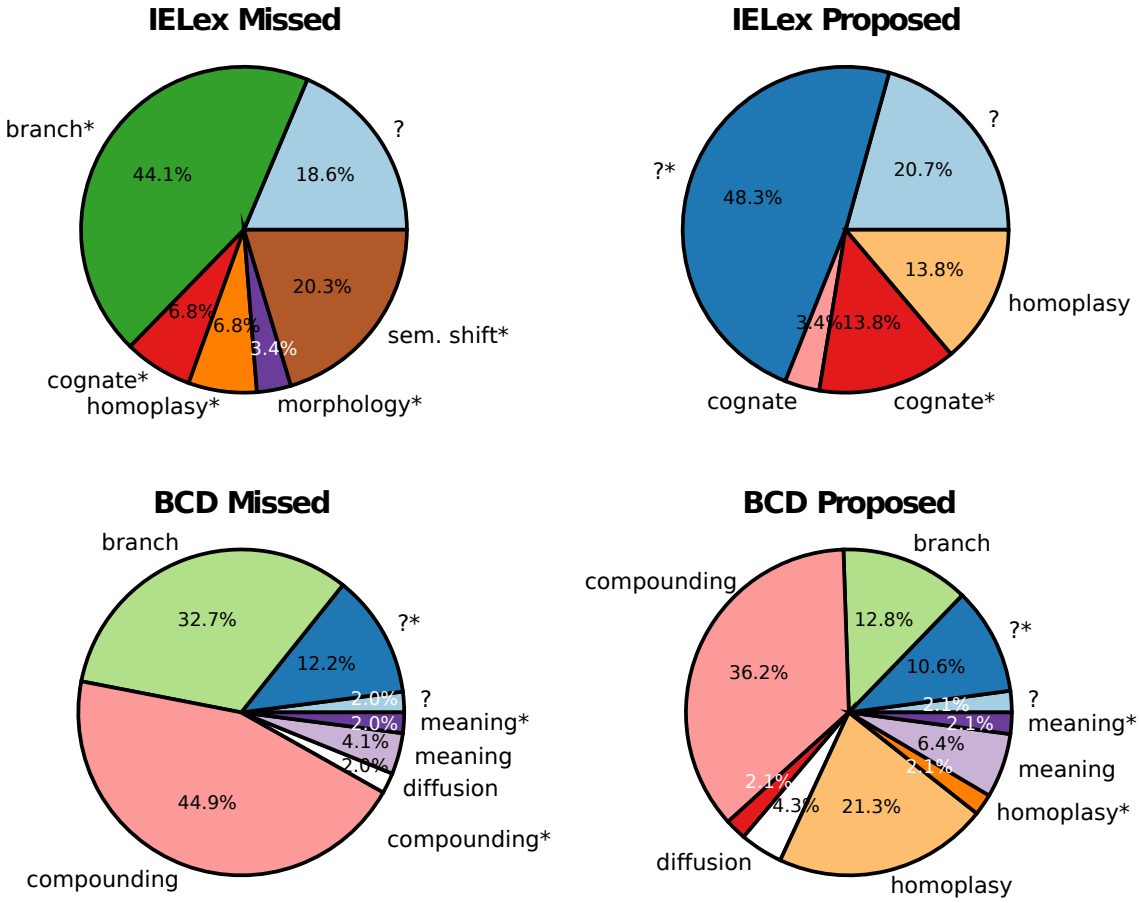

Figure 8: Detailed error analysis of the algorithmic performance on IELex and BCD. If a certain error class is followed by an asterisk, this means that we attribute the error to the gold standard rather than to the algorithm. For a detailed discussion of the different error classes mentioned in this context, please see the detailed analysis in the supplementary material.

as the PIE root *dịeu̯- 'heavenly deity, day' is a more broadly reflected candidate for the 'day' in PIE (Meier-Brügger, 2002, 187f).

Of the 29 missed cognates, the majority cannot be readily classified, as these comprise cases where a reconstruction back to the proto-language *in* the given meaning slot seems to be highly plausible. Thus, the form *kr̥-m-i- (cognate class `worm:A`) is not listed in the gold standard, but proposed by the Binary ML approach. The root is reflected in both Indo-Iranian and in Slavic (Derksen, 2008, 93f) and generally considered a good Indo-European root with the meaning 'worm, insect' (Mallory and Adams, 2006, 149). Given that 'worm' and 'insect' are frequently expressed by one polysemous concept in the languages of the world (see the CLICS database of cross-linguistic polysemies, `http://clics.lingpy.org`, List et al. 2014a), we see no reason why the form is not listed in the gold standard.

Second in frequency of the items proposed by the algorithm are cases of clear homoplasy which were interpreted as inheritance by the ML approach. As an example, consider the form *serp- (cognate class snake:E) which the algorithm proposes as candidate for the meaning 'snake'. While the cognate set contains the Latin word *serpens*, as well as reflexes in Indo-Iranian and Albanian, it may seem like a good candidate. According to Vaan (2008, 558), however, the verbal root originally meant 'to crawl', which would motivate the parallel denotation in Latin and Albanian. Instead of assuming that the noun already denoted 'snake' in PIE times, it is therefore much more likely that we are dealing with independent semantic shift.

In the case of our linguistic evaluation of the results on the Chinese data, we also find branch-specific words as one of the major reasons for the 49 forms which were proposed in the gold standard but not recognized by the best algorithm (Multi ML), but here we cannot attribute these to questionable decisions in the gold standard, but rather to the fact that many Old Chinese words are often only reflected in some of the varieties in the sample. As an example for a challenging case, consider the form 口 *kǒu* 'mouth' (cognate class mouth-Kou-222, # 31). The regular word for 'mouth' in most dialects today is 嘴 *zuǐ*, but the Mǐn dialects, the most archaic group which was the first to spread off the Sinitic family, have 喙 *huì* as an innovation, which originally meant 'beak, snout'. While *kǒu* survives in many dialects and also in Mandarin Chinese in restricted usage (compare 住口 *zhùkǒu* 'close' + 'mouth' = 'shut up)' or as part in compounds (口水 *kǒushuǐ* 'mouth' + 'water' = 'saliva'), it is only in the Yuè dialect Guǎngzhōu that it is given in the original meaning in the BCD. Whether *kǒu*, however, is a true retention in Guǎngzhōu is quite difficult to say, and comparing the data in the BCD with the more recent dataset by Liú et al. (2007), we can see that *zuǐ* is given for Guǎngzhōu instead of *kǒu*. The differences in the data are difficult to explain, and we see three possible ways to account for them: If *kǒu* was the regular term for 'mouth' in Guǎngzhōu in the data by Wang (2004), and if this term is not attested in any other dialect, we are dealing with a *retention* in the Yuè dialects, and with a later diffusion of the term *zuǐ* across many other dialect areas apart from the Mǐn dialects who all shifted the meaning of *huì*. If *kǒu* is just a variant in Guǎngzhōu as in Mandarin Chinese, we are dealing with a methodological problem of *basic word translation* and should assume that *kǒu* is completely lost in its original meaning. In both cases, however, the history of 'mouth' is a typical case of *inherited variation* in language history, namely the fact that multiple terms with a similar reference potential have been already present in the last common ancestor of the Chinese dialects, which were later individually resolved, yielding patterns that remind of *incomplete lineage sorting* in evolutionary biology (see List et al. 2016 for a closer discussion of this analogy).

The problem of inherited variation becomes even more evident when looking at the largest class of errors in both the items missed and the items proposed by the algorithm: the class of errors due to *compounding*. Compounding is a very productive morphological process in the Chinese dialects, which was favored by the shift from a predominantly

monosyllabic to a bisyllabic word structure in the history of Chinese (see Sampson 2015 and replies to the article in the same volume for a closer discussion on potential reasons for this development). This development lead to a drastic increase of bisyllabic words which is reflected in almost all dialects, affecting all parts of the lexicon. Thus, while the regular words for 'sun' and 'moon' in Ancient Chinese texts were 日 *rì* and 月 *yuè*, the majority of dialects nowadays uses 日頭 *rìtóu* (lit. 'sun-head') and 月光 *yuèguāng* (lit. 'moon-shine'). These words have been further developing in some dialect areas and yield a complex picture of denotation patterns which are extremely difficult to resolve historically. Given that we find the words also in the most archaic dialects, but *not* in ancient texts of late Hàn time and later (around 200 and 300 CE), when the supposed LCA of the Chinese dialects was spoken, it is quite difficult to explain the data in a straightforward way. We could either propose that the LCA of Chinese dialects already had created or was in the stage of creating these ancient compound words, and that written evidence was too conservative to reflect it, or we could propose that the words were created later and then diffused across the Chinese dialects. Both explanations seem plausible, as we know that spoken and written language often differed quite drastically in the history of Chinese. Comparing modern Chinese dialect data, as given in Liú et al. (2007), with dialect surveys of the late 1950s, as given in Běijīng Dàxué (1964), we can further see how quickly Mandarin Chinese words have been diffusing recently: While we find only *rìtóu*[12] as form for 'sun' in Guǎngzhōu, Liú et al. only list the Mandarin form '太陽 *tàiyáng*, and Hóu (2004), presenting data collected in the 1990s, lists both variants. We can see from these examples that the complex interaction of morphological processes like compounding and intimate language contact confront us with challenging problems and may explain why despite the automatic methods perform worst on Chinese, despite the shallow time depths of the language family.

5 Conclusion

What can we learn from these experiments? One important point is surely the striking superiority of Maximum Likelihood, outperforming both parsimony approaches. Maximum Likelihood is not only more flexible, as parameters are estimated from the data, but in some sense, it is also more realistic, as we have seen in the reconstruction of the scenario for 'eat' (see Figure 7) in the Chinese dataset, where the branch lengths, which contribute to the results of ML analyses, allow the algorithm to find the right answer. Another important point are the weaknesses of all automatic approaches and what we can learn from our detailed linguistic evaluation. Here, we can see that further research is needed to address those aspects of lexical change which are poorly handled by the algorithms. These problems include especially the problem of independent semantic shift, but also the problem of morphological change, especially in the Chinese data. List (2016) uses weighted parsimony with polarized (directional) transition penalties for multi-state characters

[12] In the Yuè dialects, this form has been reinterpreted as 'hot-head' 熱頭 *rètóu* instead of 'sun-head'.

for ancestral state reconstruction of Chinese nouns and reports an increased performance compared to unweighted parsimony. However, since morphological change and lexical replacement are clearly two distinct processes, we think it is more promising to work on the development of stochastic models which are capable of handling two or more distinct processes and may estimate transition tendencies from the data. Another major problem which needs to be addressed in future approaches is the impact of language contact on lexical change processes, as well as the possibility of language-internal variation, which may obscur tree-like divergence even if the data evolved in a perfectly tree-like manner. These instances of *incomplete lineage sorting* (List et al., 2016) became quite evident in our qualitative analysis of the Chinese and Indo-European data. Given their pervasiveness, it is likely that they also have a major impact on classical phylogenetic studies which only try to infer phylogenies from the data. As a last point, we should mention the need for increasing the quality of our test data in historical linguistics. Given the multiple questionable reconstructions we found in the test sets during our qualitative evaluation, we further think it might be fruitful, both in classical and computational historical linguistics, to intensify the efforts towards semantic and onomasiological reconstruction.

Supplementary Material

All data which was used for this study as well as all results which we produced, and the code that was used to produce these results will be published with Zenodo upon official publication of the paper. For the moment, the data is available from GitHub (anonymous GitHub-Gist, requires a GitHub-account to download):

- Austronesian: `https://gist.github.com/anonymous/06854b14b6460a42877dc9e57de9fdbd`
- Indo-European: `https://gist.github.com/anonymous/3c14548073cbdf23fe7bafa25811ccf9`
- Chinese: `https://gist.github.com/anonymous/3a51ef72e16f4ba7880751916dcfec78`

If you have problems accessing the supplementary material, please feel free to contact us.

References

Quentin D. Atkinson and Russell D. Gray. How old is the Indo-European language family? Illumination or more moths to the flame? In Peter Forster and Colin Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, pages 91–109. McDonald Institute for Archaeological Research, Cambridge and Oxford and Oakville, 2006. ISBN 9781902937335.

Mahe Ben Hamed and Feng Wang. Stuck in the forest: Trees, networks and chinese dialects. *Diachronica*, 23:29–60, 2006.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11):4224–4229, 2013.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Aalexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, Aug 2012.

Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. Beast 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4):e1003537, 04 2014. doi: 10.1371/journal.pcbi.1003537.

Claire Bowern and Quentin D. Atkinson. Computational phylogenetics of the internal structure of pama-nguyan. *Language*, 88:817–845, 2012.

Hadumod Bussmann, editor. *Routledge dictionary of language and linguistics*. Routledge, London and New York, 1996.

Běijīng Dàxué 北京大学. *Hànyǔ fāngyán cíhuì* 汉语方言词汇 *[Chinese dialect vocabularies]*. Wénzì Gǎigé, Běijing □□, 1964.

George L. Campbell and Gareth King. *Compendium of the World's Languages*, volume 1. Routledge, London and New York, 2013.

Will Chang, Chundra Cathcart, David Hall, and Andrew Garret. Ancestry-constrained phylogenetic analysis ssupport the indo-european steppe hypothesis. *Language*, 91(1):194–244, 2015.

T. Dagan, Y. Artzy-Randrup, and W. Martin. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29):10039–10044, 2008.

Rick Derksen. *Etymological dictionary of the Slavic inherited lexicon*. Brill, Leiden and Boston, 2008.

Michael Dunn. *Indo-European lexical cognacy database (IELex)*. Max Planck Institute for Psycholinguistics, Nijmegen, 2012.

Anthony W. F. Edwards and Luigi Luca Cavalli-Sforza. Reconstruction of evolutionary trees. In V. H. Heywood and J. McNeill, editors, *Phenetic and Phylogenetic Classification*, pages 67–76. Systematics Association Publisher, London, 1964.

Warren Ewens and Gregory Grant. *Statistical Methods in Bioinformatics: An Introduction*. Springer, New York, 2005.

James S. Farris. Phylogenetic analysis under dollo's law. *Systematic Biology*, 26(1):77–88, 1977.

Walter M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

Hans Geisler and Johann-Mattis List. Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics. In Heinrich Hettrich, editor, *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Reichert, Wiesbaden, 2010. Document has been submitted in 2010 and is still waiting for publication.

Russell D. Gray and Quentin D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426 (6965):435–439, 2003.

Russell D. Gray, Alexei J. Drummond, and S. J. Greenhill. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science*, 323(5913):479–483, 2009.

Simon J. Greenhill, Robert Blust, and Russell D. Gray. The austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271–283, 2008.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2015. URL `http://glottolog.org`.

H. J. Haynie and C. Bowern. Phylogenetic approach to the evolution of color term systems. *Proc. Natl. Acad. Sci. U.S.A.*, 113(48):13666–13671, Nov 2016.

Hans J. Holm. The new arboretum of Indo-European "trees". *Journal of Quantitative Linguistics*, 14(2-3):167–214, 2007.

Jīngī Hóu, editor. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 [*Phonological database of Chinese dialects]*. Shànghǎi Jiàoyù, Shànghǎi, 2004.

D. J. Hruschka, S. Branford, E. D. Smith, J. Wilkins, A. Meade, M. Pagel, and T. Bhattacharya. Detecting regular sound changes in linguistics as events of concerted evolution. *Curr. Biol.*, 25(1):1–9, Jan 2015.

Daniel H. Huson. Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

Fiona M. Jordan, Russell D. Gray, Simon J. Greenhill, and Ruth Mace. Matrilocal residence is ancestral in Austronesian societies. *Proc. R. Soc. B*, 276:1957—1964, 2009.

Gerhard Jäger. Support for linguistic macrofamilies from weighted alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 112(41):12752–12757, 2015.

Alexei Kassian, Mikhail Zhivlov, and George S. Starostin. Proto-indo-european-uralic comparison from the probabilistic point of view. *The Journal of Indo-European Studies*, 43(3-4):301–347, 2015a.

Alexei Kassian, Mikhail Zhivlov, and George S. Starostin. Proto-indo-european-uralic comparison from the probabilistic point of view. *The Journal of Indo-European Studies*, 43(3-4):301–347, 2015b.

A. Kitchen, C. Ehret, S. Assefa, and C. J. Mulligan. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. Biol. Sci.*, 276(1668):2703–2710, Aug 2009.

Guus Kroonen. *Etymological dictionary of Proto-Germanic*. Number 11 in Leiden Indo-European Etymological Dictionary Series. Brill, Leiden and Boston, 2013.

Walter J. Le Quesne. The uniquely evolved character concept and its cladistic application. *Systematic Biology*, 23(4):513–517, 1974.

Sean Lee and Toshikazu Hasegawa. Evolution of the ainu language in space and time. *PLoS ONE*, 8(4):e62243, 04 2013.

Thomas Lepage, David Bryant, Hervé Philippe, and Nicolas Lartillot. A general comparison of relaxed molecular clock models. *Molecular biology and evolution*, 24(12):2669–2680, 2007.

J.-M. List, T. Mayer, A. Terhalle, and M. Urban. *CLICS: Database of Cross-Linguistic Colexifications*. Forschungszentrum Deutscher Sprachatlas, Marburg, 2014a. Online available at `http://clics.lingpy.org`.

Johann-Mattis List. Network perspectives on chinese dialect history. *Bulletin of Chinese Linguistics*, 8:42–67, 2015.

Johann-Mattis List. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2):119–136, 2016. doi: 10.1093/jole/lzw006.

Johann-Mattis List and Robert Forkel. *LingPy. A Python library for historical linguistics*. Max Planck Institute for the Science of Human History, Jena, 2016. doi: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. URL `http://lingpy.org`.

Johann-Mattis List, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays*, 36(2):141–150, 2014b.

Johann-Mattis List, Shijulal Nelson-Sathi, William Martin, and Hans Geisler. Using phylogenetic networks to model chinese dialect history. *Language Dynamics and Change*, 4(2):222–252, 2014c.

Johann-Mattis List, Jananan Sylvestre Pathmanathan, Philippe Lopez, and Eric Bapteste. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. *Biology Direct*, 11(39):1–17, 2016.

Liú Lìlǐ 刘俐李, Wáng Hóngzhōng 王洪钟, and Bǎi Yíng 柏莹. *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí* 现代汉语方言核心词· 特征词集 *[Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]*. Fènghuáng 凤凰, Nánjīng 南京, 2007.

J. P. Mallory and D. Q. Adams. *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*. Oxford University Press, Oxford, 2006.

Michael Meier-Brügger. *Indogermanische Sprachwissenschaft [Indo-European linguistics]*. de Gruyter, Berlin and New York, 8 edition, 2002.

B. G. Mirkin, T. I. Fenner, M. Y. Galperin, and E. V. Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, 3:2, 2003.

Shijulal Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1713): 1794–1803, 2011.

Mark Pagel and Andrew Meade. BayesTraits 2.0. software distributed by the authors, November 2014.

Mark Pagel, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meade. Ultraconserved words point to deep language ancestry across eurasia. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21):8471–8476, 2013.

Helmut Rix, Martin Kümmel, Thomas Zehnder, Reiner Lipp, and Brigitte Schirmer. *LIV. Lexikon der Indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen [Lexicon of Indo-European Verbs. The roots and their primary stems]*. Reichert, Wiesbaden, 2001.

Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.

L. Sagart. *Classifying Chinese dialects/Sinitic languages on shared innovations*. CRLAO, Paris, 2011. Paper, presented at the Séminaire Sino-Tibétain du CRLAO (2011-03-28). Online available at `https://www.academia.edu/19534510`.

Laurent Sagart. Old chinese. In Simon Greenhill, editor, *Austronesian Basic Vocabulary Database*, page 331. The University of Auckland, Auckland, 2008. URL `http://language.psy.auckland.ac.nz/austronesian/language.php?id=331`.

N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406–425, 1987.

Geoffrey Sampson. A chinese phonological enigma. *Journal of Chinese Linguistics*, 43(2):679–691, 2015.

David Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

G. S. Starostin. From wordlists to proto-wordlists: reconstruction as 'optimal selection'. *Faits de langues*, 47(1):177–200, 2016. doi: 10.3726/432492_177.

Morris Swadesh. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96(4):452–463, 1952.

Morris Swadesh. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137, 1955. ISSN 00207071.

Matthias Urban. Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics*, 1(1):3–47, 2011.

Michiel Vaan. *Etymological dictionary of Latin and the other Italic languages*. Number 7 in Leiden Indo-European Etymological Dictionary Series. Brill, Leiden and Boston, 2008.

Feng Wang. Bcd: Basic words of chinese dialects, 2004.

Elias Wessen. *Die nordischen Sprachen*. de Gruyter, Berlin, 1968.

David P. Wilkins. Natural tendencies of semantic change and the search for cognates. In Mark Durie, editor, *The comparative method reviewed. Regularity and irregularity in language change*, pages 264–304. Oxford University Press, New York, 1996. ISBN 9780195066074.

Dagmar Wodtko, Britta Irslinger, and Carolin Schneider. *Nomina im Indogermanischen Lexikon [Nominals in the Indo-European lexicon]*. Winter, Heidelberg, 2008.

Ziheng Yang and Bruce Rannala. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution*, 14(7):717–724, 1997.

Kevin Zhou and Claire Bowern. Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proceedings of the Royal Society B*, 282(1815):20151278, 2015.

A Age Constraints for Indo-European

Age constraints for IELex used in Bayesian phylogenetic inference (in years before present). Prior probabilities are uniform distributions over the interval [Minimum,Maximum].

taxon/clade	minimum age	maximum age
Ancient Greek	2,300	2,500
Avestan	2,400	2,600
Classical Armenian	1,300	1,600
Gaulish	1,400	2,600
Gothic	1,600	1,700
Hittite	3,100	3,600
Latin	1,900	2,200
Luvian	3,200	3,500
Lycian	2,300	2,500
Middle Breton	400	900
Middle Cornish	400	700
Mycenaean Greek	3,000	3,500
Old Breton	1,000	1,200
Old Church Slavonic	900	1,100
Old Cornish	600	1,150
Old English	900	1,100
Old Gutnish	480	1,100
Old Swedish	380	800
Old High German	1,000	1,300
Old Irish	1,100	1,400
Old Norse	695	855
Old Persian	2,300	2,600
Old Prussian	400	600
Old Welsh	800	1,100
Oscan	2,000	2,400
Palaic	3,000	4,000
Tocharian A	1,225	1,525
Tocharian B	1,200	1,500
Umbrian	2,000	2,400
Vedic Sanskrit	2,800	3,200
Pali	200	2,300

taxon/clade	minimum age	maximum age
Balto-Slavic	1,100	3,400
Brythonic	1,500	2,500
Celtic	2,500	10,000
French/Iberian	1,200	1,600
Indic	2,200	2,900
Indo-Iranian	3,001	10,000
Iranian	2,500	10,000
Latin/Romance	1,901	2,200
Northwest Germanic	1,700	2,000
Tocharic	1,375	2,135
West Germanic	1,600	1,700

B Linguistic Evaluation

B.1 Comments on the IELex Gold Standard

In this appendix, we provide our detailed qualitative error-analysis of the results of the Binary ML Sample method compared to the Gold Standard as reflected in **IELex**. In the following table, we list two general types of errors: Forms proposed by the Gold Standard which are not proposed as root-forms by the method (section A), and forms proposed as PIE roots by the methods which are not given in the Gold Standard (section B). Our critical analysis of the forms proposed in the Gold Standard and proposed by the Binary ML Sample method was done by comparing the proposed forms and the reflexes of the cognate sets in the data with the proposals we could find in the standard literature on PIE. Thus, when dealing with predominantly Slavic and Balto-Slavic forms, we consulted Derksen (2008), when dealing with Germanic forms, we consulted Kroonen (2013), and Vaan (2008) for Latin forms. For classical examples of PIE reconstructions, we compared with Meier-Brügger (2002), for PIE nouns we compared Wodtko et al. (2008), and for verbs we compared Rix et al. (2001). In addition, we consulted Mallory and Adams (2006), since this source is useful in so far as the authors try to provide a classical onomasiological reconstruction by comparing the distribution of cognate sets, and their meaning in the descendant languages. Where available, we also compared the data with the database of *Cross-Linguistic Colexifications* (CLICS, `http://clics.lingpy.org`, List et al. 2014a), which lists frequently recurring polysemy relations for more than 220 different languages and 1280 different concepts. Based on this comparison, we try to identify the sources of the error (column ES in the table) as either originating from the Gold Standard (GS) or from the automatic method (ASR). We further classify general *causes* of errors (column EC in the table) by distinguishing *branch-specific cognates* which are erroneously reconstructed back to the root (B in column EC), obvious instances of *semantic shift* which allow us to assume that the form had another meaning in PIE (S in column EC), erroneous or dubious *cognate judgments* (C in column EC), parallel semantic shift (H in column EC, meaning *homoplasy*), and problematic representations of complex *morphology* (M in column EC). We further clarify our decisions in a note accompanying each of the errors.

B.1.1 Words proposed by the gold standard but not by the algorithm

#	CC	PIE FORM	ES	EC	NOTE
1	`at:R`	*$h_2$ed	GS	B	Doubtful root which is only reflected in Germanic and Romance with a rather vague meaning.
2	`bad:U`	*uba	GS	B	IELex mentions itself that this root is only attested in Germanic languages.
3	`black:B`	*s̥uordos	GS	B	This root is only reflected in Germanic, and the reconstruction back to PIE is based on the proposed cognacy with Latin *sordes* 'dirt'. Assuming that the word had the same meaning in PIE is by no means straightforward.

4	blood:I	*kreu̯h$_2$-	ASR	?	This seems to be a clear-cut example of an algorithmic failure, as the root is reflected in the meaning in Slavic, and in only slightly derived meanings in Latin and Sanskrit. Here, we can see that ASR algorithms, being deprived of additional information on words with similar meanings, can only reconstruct what they find in the narrow windows of concept slots.
5	blow:A	*b^{h}leh$_1$-, *b^{h}l̥h$_1$-	GS	S	This root is reconstructed back to Indo-European, but not necessarily in this meaning, as Rix et al. (2001: 87) give 'to cry' ('heulen') as the reconstructed meaning.
6	breathe:P	*pneu̯-	GS	B	Highly questionable root, as it only occurs in Ancient Greek and has no reflexes in any other language family. It is also flagged as questionable in the Lexikon Indogermanischer Verben (Rix et al. 2001: 489) and should be discarded.
7	burn:O	*su̯el-	GS	B	Not to be reconstructed in this meaning, as Mallory and Adams (2006: 89) also only reconstruct the word *d^{h}eg^{wh}- 'to burn'. Rix et al. (2001: 609) reconstruct this in the meaning of 'schwelen, brennen' ('to smolder'), but they only list Lithuanian and Germanic examples.
8	burn:U	*h$_1$eu̯s-	GS	S	Apparently reflected in Latin, Sanskrit, and Greek, but not consistently with the same meaning.
9	cold:B	*kalda-	GS	B	The IELex mentions itself that the root has no cognates outside the Germanic branch.
10	day:B	*d^{h}ōĝ^h-e/os-	GS	S	Kroonen (2013: 86f) confirms the reconstruction of the root in Proto-Germanic and Indo-Iranian. The meaning, however, is by no means clear, as the PIE root *di̯eu̯- 'heavenly deity, day' is a more broadly reflected candidate for the meaning 'day' in PIE.
11	drink:H	*h$_1$eg^{wh}-	GS	B	Given that this root is only reflected in Tocharian and Anatolian, it is by no means clear whether it reflects a PIE root. Mallory and Adams (2006) also do not reconstruct this root, but prefer *peh$_3$(i)- 'to drink', which is much more widely reflected.
12	dry:D	*tr̥s-ú-, *tērs-o-	GS	B	This is in clear contradiction with Mallory and Adams (2006), who list PIE *saus- 'dry'. The root is also only reflected in Albanian and Germanic.
13	dust:H	*prs-o-	GS	B	This root occurs only in Balto-Slavic, as already indicated in IELex itself.
14	far:G	*u̯i-itós	GS	B	This root occurs only in Germanic.
15	fat:U	*smer-u-	GS	B	This root occurs only in Germanic.
16	father:B	*atta-	GS	H	This root is labelled as an onomatopoetic word in IELex, so it reflects a parallel development rather than a valid word for PIE.

17	fire:E	*h_1ng^w-ni-	ASR	?	Mallory and Adams only list PIE *$péh_2ur$ 'fire' for the concept 'fire', but given that this word occurs as reflex in Slavic, Sanskrit, and Latin without any change in meaning, it is straightforward to reconstruct it back to PIE. The problem remains for IE reconstruction, that one has to explain why there were two different words for fire, namely *$péh_2ur$ and *h_1ng^w-ni- (see discussion in Wodtko et al. 2008: 540-545).
18	fire:J	*h_2eh_1-t(e)r-	GS	B	Reconstructing this word in the meaning 'fire' back to PIE is not justified, given that it only occurs in Indo-Iranian, also in shifted meanings.
19	fish:B	*pisk-, *peisk-	GS	B	Mallory and Adams (2006) reconstruct only *$d^h\hat{g}^huh_2$- in this meaning, as the form *pisk- is only reflected in Romance and Germanic.
20	flow:I	*sreu̯-	ASR	?	Rix et al. (2001: 588) reconstruct this root as 'strömen, fließen' ('to stream, to flow'.
21	fog:T	*(s)neu̯d^h-, *(s)nou̯d^h-	GS	B	Only reflected in Avestan.
22	good:Ae	*b^hedró-	GS	M	This is the suppletive stem for 'good' in the Germanic languages, and apparently also reflected in Vedic Sanskrit. Reconstructing it as a root itself does not seem to be justified, especially since the normal root for 'good' in PIE would be the widely reflected *h_1(e)su- (see Mallory and Adams 2006).
23	hair:Q	*pulo-	GS	S	Mallory and Adams (2006: 97) reconstruct *$\hat{k}$ripo- for 'hair', and the word shows a much better distribution, as it is only reflected in Romance, and in a shifted meaning in Vedic Sanskrit. Given the multiple possibilities for semantic S involving 'hair', as it is also reflected in the CLICS database (List et al. 2014), it is much more likely that this is an instance of semantic S of a word that meant something different in PIE.
24	hand:E	*mon-u-	GS	S	No reason to assume that this Romance cognate meant anything close to 'hand' in PIE, where the more common root *$\hat{g}^hes$-r- would be the best candidate for this meaning (Mallory and Adams 2006).
25	hold:M	*se$\hat{g}^h$-	GS	C	Cognacy with Portuguese as the only reflex of this root apart from Old Greek is quite doubtful. Rix et al. (2001: 515f) further reconstruct a different meaning for the term, 'in den Griff bekommen' ('to get hold off') rather than 'to hold'.
26	I:D	*me-	GS	M	Complex paradigm for 'I' in PIE. Handling this as two cognate sets is not straightforward in the Gold Standard, as the paradigm developed in one set, rather than independently.

27	know:A	*ĝneh$_3$-	GS	S	A clear case of parallel semantic S of an extremely well-attested root in PIE. The common term for 'to know' in PIE was *weid- (Mallory and Adams 2006), and this root meant something like 'to recognize' rather than 'to know'.
28	lake:L	*h$_2$ep-	GS	S	There is no clear-cut reason to assume that this root, which is reflected in many different languages, meant exactly 'lake' in PIE (see Mallory and Adams 2006, 127).
29	laugh:D	*ĝelh$_2$-	GS	B	Only reflected in Armenian and Greek in this meaning, and given the supposed closeness of both branches, it is by no means straightforward to reconstruct it back in exactly this meaning into PIE.
30	laugh:H	*smei̯-	GS	B	This root is only reflected in the Balto-Slavic branch, and in the meaning 'to smile' in Sanskrit smáyate (Derksen 2008: 456). Following Mallory and Adams (2006: 359f), it is furthermore much more likely that the original word for 'to laugh' in PIE was *k^{h}a-, with reflexes in Latin, Sanskrit, and Greek.
31	leg:P	*kokso-	GS	B	Only reflected in Celtic languages, a much better candidate for 'leg' in PIE would be *sókwt- (Mallory and Adams 2006: 182f) with reflexes in Hittite and Sanskrit.
32	lie:A	*leg^h-	ASR	?	Mallory and Adams reconstruct *ḱei- for 'to lie' (Old Greek κεῖμαι, Sanskrit śāyayati 'to lay down', see Pokorny 1959), but given how widely this candidate is reflected, it is at least equally good for the meaning 'to lie' as the one proposed by Mallory and Adams.
33	man:F	*h$_2$nḗr	ASR	?	Mallory and Adams reconstruct the same root, and given its wide distribution in the meaning of 'man' it seems the best candidate for PIE.
34	not:F	*meh$_2$	GS	S	As IELex notes, the root reconstructed to PIE is not the regular term for 'not', as it is used in specific grammatic constructions (as, for example, in Old Greek).
35	one:A	*(H)ói̯nos, *(H)ói̯kos, *(H)ói̯u̯os	ASR	?	This is a regular candidate for the word of 'one' in PIE, as also given in Mallory and Adams, who give the form *h□oin- (97).
36	river:O	*h$_2$ek^w-eh$_2$-	GS	S	Form in GS meant 'water' in PIE. Although a shift from 'water' to 'river' is likely according to CLICS (List et al. 2014), this meaning is an innovation in Germanic.
37	rub:L	*melh$_1$-	GS	C	Form in GS is not reflected in the standard literature (LIV and LIN).
38	scratch:B	*gerb^h-	GS	B	Form in GS is only reflected in few Germanic languages, probably with a wrong cognate assignment.

39	see:D	*derḱ-	GS	S	Form is also reflected as the valid form for ‘to see’ in Mallory and Adams (2006: 98), but it is not clear whether ‘to see’ reflects the original meaning, as Rix et al. (2001, 122) reconstruct it as ‘to catch sight of’. Given the strong polysemy relations between verbs for ‘to look’, ‘to see’, and ‘to find’, as reflected in CLICS (List et al. 2014), it is quite likely that this form originally reflected another meaning, especially given that the form *u̯ei̯d- is a widely reflected good candidate for ‘to see’ in PIE.
40	short:M	*mr̥ǵʰ-ú-	ASR	?	Form is only reflected in Celtic, Romance, and Greek, and in shifted form in Germanic, but Mallory and Adams (2006: 317) also reconstruct the same meaning.
41	sit:B	*sed-	ASR	?	Clear root for PIE, widely attested in Slavic, Romance, Germanic, and Sanskrit.
42	skin:B	*pel-ni-	ASR	?	Form in GS is a good PIE root, but not necessarily with the meaning ‘skin’, as the meaning of the reflexes differs greatly. The GSR form derives from a PIE verb meaning ‘to cover’, so independent semantic S is likewise possible. Mallory and Adams (2006), however, also propose *péln- for ‘skin’ in PIE.
43	sleep:E	*drem-	GS	B	This form is only reflected in Romance and Vedic Sanskrit, while the form *swep- has a much better distribution (Mallory and Adams 2006).
44	small:H	*mei̯-	GS	C	Wrong cognate judgments in the database, since neither Russian *malenkij* nor English *small* go back to this root (see Derksen 2008: 308).
45	snake:D	*h$_2$enguu̯i	ASR	?	Mallory and Adams reconstruct *h$_1$ógʷʰis, but they list the same reflexes for the word and propose this as the basic word for the meaning ‘snake’.
46	snow:B	*sneiǵʷʰ-	ASR	?	It is surprising that the algorithm missed this root, which is one of the classical roots which was recognized very early in the history of PIE. Yet remembering that the root is not necessarily reconstructed as ‘snow’ in PIE, but, given its verbal reflexes in Vedic Sanskrit, as ‘to be sticky’ (Meier-Brügger 2002: 173f), we may as well deal with an independent semantic shift in some PIE languages, as the root itself occurs only in Romance, Slavic, and Germanic in the meaning ‘snow’.
47	think:B	*tong-	GS	B	Root only reflected in Germanic languages with spurious reflexes in semantically shifted form in other branches. A better candidate for PIE would be *men- “the mind or to think”.
48	this:G	*kos, *koh$_2$, *kod	GS	H	Given the complex structure of pronouns in PIE, it is not straightforward, to only reconstruct this form in the meaning of ‘this’ back to PIE, as also reflected in Mallory and Adams (2006), who propose the widely reflected form *so- instead.

49	tooth:C	*ĝembʰ-	GS	H	There is no reason to assume that this word meant 'tooth' in PIE, given that there is the excellent candidate *h$_1$dónt- (Mallory and Adams 2006). The original meaning of this word must have been something else, so that its distribution across some PIE languages can be explained.
50	vomit:S	*h$_1$reug-	GS	H	No need to reconstruct this form back to PIE, since it is only reflected in two languages of Romance and Serbo-Croatian. It is more likely that this is independent semantic semantic shift (original meaning 'to throw out'), if the words are cognate at all.
51	walk:G	*gʰredʰ-	GS	B	Not clear why this form is reconstructed as the PIE root, given that it is only reflected in Germanic in this meaning. More likely is the root *h$_1$ei-, as suggested by Mallory and Adams (2006), as it has more reflexes in PIE.
52	warm:D	*tep-	GS	B	Mallory and Adams (2006) propose *gʷʰermós for 'hot'. This root has a much wider distribution, while *tep is mostly reflected in Slavic and in derived variants in Latin and Sanskrit (Derksen 2008: 496). The cognates with Celtic in IELex seem doubtful.
53	wash:C	*leh$_2$u̯-	GS	C	Wrong cognate assignment in the source since Romance and Albanian reflexes are not annotated.
54	water:A	*h$_2$ekʷ-eh$_2$-	GS	B	Mallory and Adams (2006) list *wódr̥ as the root for 'water' in PIE, and this seems to be a much better candidate than this one, which is only in Romance reflected in this meaning, and may have easily shifted from another meaning, if one reconstructs it back to PIE.
55	water:E	*h$_2$ep-	GS	S	As mentioned above, this root is also reconstructed as 'lake' in IELex, but it is more likely, following Mallory and Adams (2006: 127) that the word meant 'body of water' originally and then shifted into the meaning.
56	wet:I	*u̯ed	GS	B	Semantic change from 'water' to 'wet' is likely according to CLICS, but it is not clear why this should have already happened in PIE times.
57	white:E	*h$_2$elbʰós	GS	B	The IELex form is only reflected in Romance in this meaning and as meaning 'cloud' in Hittite.
58	worm:B	*u̯r̥mi-	GS	B	The IELex form is only reflected in Germanic and Romance
59	year:B	*i̯eHr-	GS	S	Mallory and Adams (2006: 300) reconstruct a specified meaning for this root, namely 'season', rather than 'year', which already has a good candidate with *wet-.

B.1.2 Words proposed by the algorithm but not by the gold standard

#	CC	PIE FORM	ES	EC	NOTE
1	bad:Ac	Ø	ASR	?	Only in Hittite, Luwian, and Tocharian.
2	black:P	Ø	ASR	?	Only in Anatolian.
3	breathe:O	*h$_2$énh$_1$mi	GS	?	Reflected in Germanic and Sanskrit in the data, and also reconstructed in this meaning by Mallory and Adams (2006: 189).
4	cold:A	Ø	GS	C	Usually, and Indian form, but IELex also lists cognates in the Baltic languages, which do not seem very likely.
5	dry:F	*h$_1$so̯us-o-	GS	?	Mallory and Adams (2006) list *sau̯s- as the PIE root for the meaning 'dry', which finds regular reflexes in a widre range of languages and sub-branches.
6	dust:Q	*d^{h}uns-to-	GS	C	The root is spuriously reflected in Germanic, Celtic, Tocharian, and Lituanian. It is quite likely that we are dealing with wrong cognate assessments here, as the forms in Celtic are obviously borrowed, and the Baltic forms are obscure. The Germanic form goes probably back ot a PIE form, but in another meaning (see Kroonen 2013: 109).
7	far:B	*per-n-oi̯	ASR	C	The reflexes all represent partial cognates of the PIE root *per, which is reflected in the Germanic reflexes (Kroonen 2013: 137), but also in other branches. The inability to handle partial cognacy here leads to the algorithmic error.
8	fat:X	*poi, pī-	ASR	?	Root only reflected in Old Greek, Avestan, and Old Prussian.
9	flow:H	*tek-	GS	C	Root *tek- is a classical example for Slavic, perhaps going back to PIE (Derksen 2008: 489f), but in a slightly shifted meaning, but IELex also lists unrelated forms, namely reflexes of Germanic *rinnan (Kroonen 2013: 413), which are completely unrelated.
10	fog:A	*h$_3$mig^h-leh$_1$	ASR	H	Root is reflected in Proto-Slavic and can be reconstructed to PIE (Derksen 2008:338f), but here it probably reflected another meaning, such as 'soft rain' (Mallory and Adams 2006: 127).
11	hair:J	*u̯ol-o-	GS	?	Root is reflected in Slavic and Indo-Iranian, also with similar meanings (Derksen 2008: 526), and IELex lists also Celtic cognates. The alternative word for 'hair' *pel- (see hair:Q) is also not an extremely good candidate.
12	hold:E	*der-	GS	?	IELex lists Slavic and Indo-Iranian cognates for this cognate set, which may well go back to PIE, in a meaning, also very close to 'to hold' (Derksen 2008: 137f).
13	lake:E	Ø	GS	?	Root is only reflected in Celtic and Romance, and as a borrowing in English 'lake', so no reason to assume that it goes back to PIE (not to even speak of keeping the same meaning there).

14	laugh:A	*k^{h}a-	GS	?	Mallory and Adams propose this root for the meaning ‘to laugh’ in PIE.
15	river:L	*h$_2$ep-	GS	?	The form is reflected across multiple branches and may be a much better candidate than *h2ekweh2 as proposed as PIE word for ‘river’ in IELex.
16	rub:H	*terh$_1$-	GS	?	Form is reflected in the meaning ‘to rub, to bore’ across Slavic and Greek.
17	scratch:T	*(s)k^{w}er-	GS	?	IELex lists the PIE root for this word, but in a note rather than in a specific field for the PIE language. The etymology of the words listed in the cognate sets itself, however, is not necessarily clear, as the word shows reflexes in Slavic, which are, however, not among the traditionally acknowledge roots for Proto-Slavic.
18	short:D	*(s)ker-	ASR	H	The form may be an independent innovation in both Slavic and Germanic.
19	skin:E	Ø	GS	C	Wrong cognate assignment between reflexes of Slavic *skorà- (Derksen 2008: 452) and Germanic hūdi- (Kroonen 2013: 251f). This seems to be a technical error, as it seems unlikely that a human would come up by assuming cognacy for the two words.
20	small:A	Ø	ASR	?	Not clear why the algorithm would reconstruct the forms back to the root, as they only occur in Hittite and Waziri. This reflects another problem of automated ASR: The algorithms lack the knowledge regarding the importance of the evidence.
21	snake:E	*serp-	ASR	H	The cognate set contains the Latin word *serpens* and its reflexes in Indo-Iranian languages and Albanian. According to Vaan (2008: 558), the verbal root meant ‘to crawl’, which would motivate the denotation in Latin and Albanian. Instead of assuming that the noun already denoted ‘snake’ in PIE times, it is, however, much more likely that we are dealing with independent semantic shift.
22	snow:D	*ĝ^{h}éi-mn̥-	ASR	H	The form has probably independently shifted from the original meaning ‘frost, cold’, which is a very likely shift according to CLICS.
23	think:S	Ø	ASR	?	Cognate class only reflected in Indo-Iranian.
24	this:B	*so (*to)	GS	?	This form is also preferred by Mallory and Adams (2006).
25	walk:S	*h$_1$ei̯-	GS	?	This is the form also proposed by Mallory and Adams (2006).
26	water:B	*u̯ódr̥	GS	?	This form is also recommended by Mallory and Adams (2006).
27	wet:K	Ø	ASR	?	This form is only reflected in Greek and Latin, and independent semantic shift is as likely as an unrecognized instance of borrowings.
28	white:L	*h$_2$erĝ-	GS	?	This form surely reflects a PIE root (Wodtko et al. 2008: 317-322), and the meaning can also be closely reconstructed as ‘white, shining’. Given the word’s distribution in the most ancient languages of PIE further supports the reconstruction.

29	worm:A	*kr̥-m-i-	GS	?	Mallory and Adams (2006: 149) gloss the root as 'worm, insect', and Derksen (2008: 93f) gives the same reflexes for this root which is reflected in Indo-Iranian and in Slavic.

B.2 Comments on the BCD Gold Standard

In the following tables we list detailed comments regarding the differences between the MultiML analysis and the gold standard as given in the **BCD** for the Chinese dialects. The structure of the tables is identical with the structure we gave for the comment on the data in IELex, but our codes for the classes of errors (column EC) have changed, and we distinguish now between errors due to the misinterpretation of *branches* in the data (B), *meanings* in the sample which are hard to translate or erroneously translated (M), patterns of *compounding* (C) which cannot be handled by the methods, cases of *homoplasy* (H), and instances of *diffusion* (D). Instead of proto-forms, we list Chinese characters (column CF). Our comments draw from an inspection of alternative resources, like the dataset by Liú et al. (2007), and the alternative interpretation of Old Chinese translations provided in Sagart (2008), and by inspecting corpora of Classical Chinese texts (as provided by the *Chinese Text Project*, `http://ctext.org/`) in order to check the Old Chinese translations proposed for the concepts in the gold standard.

B.2.1 Words proposed by the Gold Standard but not by the algorithm

#	CC	CF	ES	EC	NOTE
1	back-Bei-1082	背	ASR	B	See note on 背脊 *bèijǐ* below.
2	because-Wei-69	為	Ø	M	A bad concept for reconstruction which would better be excluded, as causal relations are not necessarily expressed with conjunctions.
3	belly-Fu-1124	腹	ASR	C	See note on 肚 *dù* below.
4	bite-Nie-309	啮	ASR	C	See note on 咬 *yǎo* below.
5	black-Hei-1349	黑	ASR	B	Not at all clear how the algorithm could miss this word.
6	cold-Han-458	寒	GS	?	See note on 冷 below.
7	correct-Shi-679	是	GS	?	It seems unlikely that this is the correct word for 'correct' in Old Chinese.
8	cut-Zhuo-664	斫	GS	?	Sagart (2008) does not propose this word for 'to cut', and it is not clear which is the right form for Old Chinese.
9	day-Tian-373	天	GS	?	Sagart (2008) gives 日 *rì* which seems to be the better choice for translation.
10	dog-Quan-875	犬	ASR	B	Given the distribution of 狗 *gǒu* during Hàn times (around 100 CE), it seems likely that this word was not the only form present in Old Chinese.
11	drink-Yin-1328	飲	ASR	B	Due to an innovation in the Mǐn dialects, the algorithm has difficulties to uncover this as the ancient form.

12	dry-Zao-850	燥	GS	?	Sagart (2008) gives 乾 as the correct form.
13	dull-Yu-552	愚	GS	M	See note on 笨 *bèn* below.
14	ear-Er-1058	耳	ASR	C	See explanation to 耳朵 *ěrduō* below.
15	eye-Mu-926	目	ASR	C	See explanation to 目珠 *mùzhū* below.
16	fall-Luo-1161	落	ASR	B	Difficult to explain why the algorithm failed here, as the word is reflected in the ancient dialect groups.
17	fear-Wei-913	畏	ASR	B	See note to 怕 *pà* below.
18	float-Piao-823	漂	ASR	B	Word is difficult to reconstruct due to its distribution.
19	freeze-Dong-179	凍	ASR	C	See note to 結冰 *jiébīng* below.
20	fruit-Guo-715	果	ASR	C	See note to 水果 *shuǐguǒ* below.
21	give-Yu-64	與	ASR	B	Word only reflected in one Mĭn dialect in this form, therefore difficult for the algorithm to identify it as the retention.
22	if-Ruo-1152	若	Ø	M	A bad concept for reconstruction which would better be excluded, as conditions were and are not necessarily overtly marked in Chinese.
23	knee-Xi-1134	膝	ASR	C	See note to 膝頭 *xītóu* below.
24	leaf-Xie-236	葉	ASR	C	See note to 葉子 *yèzī* below.
25	leftside-Zuo-491	左	ASR	C	See note to 左邊 *zuǒbiān* below.
26	leg-Jiao-1101	腳	ASR	B	There are problems with the concept, as there are multiple transitions between 'leg' and 'foot' in Chinese dialects.
27	live-Sheng-897	生	ASR	B	Only reflected in Guǎngzhōu, therefore difficult to identify as retained form.
28	louse-Shi-1171	虱	ASR	C	See note to 虱子 *shīzī* below.
29	man-Nan-902	男	ASR	C	See note to 男個 *nángè* below.
30	moon-Yue-690	月	ASR	C	See note on 月光 *yuèguāng* below.
31	mouth-Kou-222	口	ASR	B	Word only occurs in Guǎngzhōu and it is therefore hard to reconstruct it.
32	name-Ming-249	名	ASR	C	See note on 名字 *míngzì* below.
33	night-Ye-361	夜	ASR	C	See note on 夜裡 *yèlǐ* below.
34	nose-Bi-1353	鼻	ASR	C	See note on 鼻頭 *bítóu* below.
35	push-Tui-637	推	ASR	?	Clear failure of the algorithm.
36	red-Chi-1215	赤	ASR	D	See note on 紅 *hońg* below.
37	rightside-You-226	右	ASR	C	See note on 左邊 *zuǒbiān* below.
38	rope-Sheng-1016	繩	ASR	C	See note on 繩子 *shéngzī* below.

39	small-Xiao-461	小	ASR	B	Clear failure of the algorithm.
40	smell-Xiu-320	嗅	ASR	B	Clear failure of the algorithm.
41	stand-Li-979	立	ASR	B	Difficult to reconstruct, as not many dialects retain the form.
42	stone-Shi-967	石	ASR	C	See note to 石頭 *shítóu* below.
43	suck-Xi-257	吸	GS	?	See note on 吮 *shǔn* below.
44	sun-Ri-668	日	ASR	C	See note on 日頭 *rìtóu* below.
45	tongue-She-1140	舌	ASR	C	See note on 舌頭 *shétóu* below.
46	tooth-Ya-869	牙	ASR	C	See note on 牙齒 *yáchǐ* below.
47	who-Shui-1208	誰	ASR	B	Difficult to reconstruct, as not many dialects retain the form.
48	woman-Nu-399	女	ASR	C	See note to 女個 *nǚgè* below.
49	you-Ru-766	汝	ASR	B	Clear failure of the algorith, but difficult, due to sparse reflexes.

B.2.2 Words proposed by the algorithm but not by the gold standard

#	CC	CF	ES	EC	NOTE
1	back-BeiJi-1086	背脊	GS	C	Almost all dialects in the data show bisyllabic forms for ‘back’, while we assume the monosyllabic form 背 *bèi* for Old Chinese. This may, however, well be a sampling error, as the more recent data by Liú et al. (2007), for example, lists only sporadically bisyllabic forms for some of the Mandarin dialects.
2	because-YinWei-334	因為	Ø	M	It is extremely difficult to identify the exact words that were used for this meaning in Classical Chinese, as causal sentences are differently constructed in the language. It seems to make more sense to exclude the word completely from lexicostatistical datasets of Chinese dialects.
3	belly-Du-1069	肚	ASR	C	The original Classical Chinese word for ‘belly’, 腹 *fú* is still reflected in the ancient Mǐn dialects, but has innovated in all other dialect groups. The Mǐn dialects, however, retain the word in compounds, which makes it impossible to reconstruct the word with methods that do not take partial cognacy into account. Reconstructing 肚 for ‘belly’, however, is wrong, as the word only occurs in Post-Hàn times (after 200 CE) in the ancient literature.

4	bite-Yao-272	咬	ASR	?	The gold standard lists 嗞 *niè* as the basic word for 'to bite' in Ancient Chinese, Sagart (2008), on the other hand, lists 噬 *shì*. Both forms are possible words for 'to bite', occuring in early sources of Hàn times. The word 嗞 *niè* is still reflected in the Hakka dialects in the sample, and Liú et al. (2007) give the same word for Měixiàn Hakka. Nowadays, however, most dialects have innovated the word to the form 咬 *yǎo* which is also reflected in all Mǐn dialects. If the word form in the Hakka dialects indeed reflects the ancient Chinese character, this would suggest that the modern word for bite, which occurs first in post-Hàn sources (around 300 CE) of ancient texts, spread across all dialects after the separation of Hakka. An alternative scenario might be an innovation in the Hakka dialects which is now erroneously treated as a reflex from Ancient Chinese 嗞 *niè*.
5	black-Wu-79	烏	ASR	H	This form is a clear innovation in some dialects and was probably reconstructed by the algorithm due its presence across multiple dialects. From Old Chinese sources, however, we know that the original word for 'black' was 黑 *hēi*.
6	cold-Leng-178	冷	GS	?	The gold standard only lists 寒 *hán* as possible word for 'cold'. However, the word 冷 *lěng* also occurs in early Hàn texts.
7	correct-Zhao-944	對	ASR	H	The form is probably an independent innovation, but it may also have been assembled due to wrong sampling, as all dialects in Liú et al. (2007) only show the form 對 *duì*. Sagart (2008) further proposes 正 *zhèng* as the original word for 'correct'. This may, however, also be due to recent borrowings, as the meaning of this word in ancient texts was almost exclusively 'to answer'.
8	cut-Zhan-663	斬	ASR	M	The form 斬 *zhǎn* can be found in the meaning 'to cut' in texts from Hàn-time. The concept 'to cut', however, is notoriously difficult, especially in Chinese, as the meaning is not sufficiently specified. As a result, scholars differ regarding their preferred form, and while the gold standard lists 斫 *zhé*, Sagart (2008) gives 斷 *duàn*.
9	day-Ri-668	日	GS	?	The gold standard lists 天 *tiān* for 'day', while Sagart (2008) gives 日 *rì*. While 天 *tiān* originally means 'sky', 日 *rì* originally means 'sun'. That the latter was used already early to denote the 'day' (opposed to the 'night') is reflected in ancient texts.
10	dog-Gou-877	狗	GS	?	Although 犬 *quǎn* is the earliest Chinese word for dog, 狗 *gǒu* also occurs rather early, although the meaning seems to be more specific.
11	drink-Chi-243	吃	ASR	H	This is a semantic shift from the word 吃 *chī* for 'to eat' to 'to drink' in some dialects.
12	dry-Gan-505	乾	GS	?	Sagart (2009) lists the form 乾 *gān* as the regular form to denote 'dry' in Old Chinese.

13	dull-Ben-982	笨	GS	M	笨 *bèn* ‘stupid’ reflects a mistranslation in the gold standard, as ‘stupid’ is clearly not a basic word, opposed to ‘dull’.
14	ear-ErDuo-1064	耳朵	ASR	B	The older form was clearly monosyllabic 耳 *ěr*, yet the compound form 耳朵 *ěrduō* is reflected across many dialects, and not restricted ot the Mandarin group alone. The fact that the Mǐn dialects show innovations, however, confirms that the compound form must be secondary.
15	eye-MuZhu-927	目珠	ASR	H	The Mǐn dialects come closest to the ancient form 目 *mù* for ‘eye’, yet they show an innovation in the compound form 目珠 *mùzhū* ‘eye-pearl’ which is also reflected in dialects of the Hakka group. This is most likely an independent innovation, as the motivation is rather clear-cut. It shows, however, how difficult it is to distinguish between the likelihood of independent motivations for compound word creation and inherited forms.
16	fall-Die-1219	跌	GS	H	The word 跌 *diē* os rather an (independent) innovation in some of the dialects than a retention in the meaning ‘to fall’. Its original meaning is ‘to stumble’, rather than ‘to fall’.
17	fear-Pa-539	怕	ASR	D	The classical word for ‘to fear’ is 畏 *wèi*. This is only reflected in Maixian Hakka and therefore very difficult to reconstruct.
18	freeze-JieBing-1008	結冰	ASR	C	The verb-noun compound 結冰 *jiébīng* is reflected in most of the dialect varieties. However, since compound verbs of this form were not present in Old Chinese, the original form 凍 *dòng* is hard to reconstruct.
19	fruit-ShuiGuo-716	水果	ASR	C	The compound 水果 *shǔiguǒ* is reflected in a multitude of dialects.
20	give-Fen-188	分	ASR	H	The form 分 *fēn* for ‘to fear’ is clearly an innovated form.
21	if-YaoShi-1187	要是	Ø	M	It is extremely difficult to identify the exact words that were used for this meaning in Classical Chinese, as conditional sentences are not necessarily overtly constructed. It seems to make more sense to exclude the word completely from lexicostatistical datasets of Chinese dialects.
22	knee-XiTou-1135	膝頭	ASR	C	膝頭 *xītóu* is a frequently occurring word in the Chinese dialects, but as with other compound words, like for ‘stone’, ‘moon’, ‘sun’, it is difficult to say, when compounding exactly started, and whether this word would qualify as an ancestral form to all dialects.
23	leaf-XieZi-240	葉子	ASR	C	葉子 *yèzǐ* is a frequently occurring word in the Chinese dialects, but as with other compound words, like for ‘stone’, ‘moon’, ‘sun’, it is difficult to say, when compounding exactly started, and whether this word would qualify as an ancestral form to all dialects.

24	leftside-ZuoBian-496	左邊	ASR	C	This is a clear example for compounding problems and the difficulties to handle them in phylogenetic analyses using only gain and loss models. The first element of the compound 左邊 *zuǒbiān* 'left side' is clearly a retention from Ancient Chinese. However, the second element was later introduced in a wide range of languages who went through stages of disyllabification in the history of Chinese and therefore added new element to monosyllabic words (especially nouns and verbs). Without the proper external knowledge, this phenomenon is very difficult to handle.
25	leg-Tui-1128	腿	ASR	B	腿 *tuǐ* is a clear later innovation.
26	live(alive)-Huo-793	活	ASR	B	活 *huó* seems like a clear later innovation, as the original word was 生 *shēng*.
27	louse-ShiMu-1175	虱母	ASR	C	Not likely to be the original word, given that Old Chinese tended to be monosyllabic.
28	man-NanGe-903	男個	ASR	C	男個 *nángè* is a rather rare form of the word for 'man'. Given the tendency of monosyllabicity, this form is also not likely to have been used to denote 'man' in the ancestor of the Chinese dialects.
29	moon-YueGuang-693	月光	ASR	C	月光 *yuèguāng* is a rather pervasive innovation which is reflected in very different Chinese dialect varieties (see also scenario in List 2016: 132).
30	mouth-Hui-315	喙	ASR	B	喙 *huì* is an innovation in the Mǐn dialects (the word originally meant 'beak'). However, the real Old Chinese word for 'mouth', 口 *kǒu* has been replaced in all dialects and can therefore not be reconstructed (yet it survives as measure word and in abstract constructions).
31	name-MingZi-251	名字	ASR	C	名字 *míngzì* is the normal form in most dialects. This reflects general questions of compounding, as also illustrated in the comments to 'stone', 'sun', 'moon', etc.
32	night-YeLi-367	夜裡	ASR	H	Words for 'night' find many different expressions in the Chinese dialects, and it is difficult to trace their origin. 夜裡 *yèlǐ*, however, is a clearly secondary reconstruction.
33	nose-BiGong-1355	鼻公	ASR	C	鼻公 *bígōng* reflects a clear later innovation.
34	push-Sang-644	搡	ASR	B	搡 *sāng* is not likely to be an ancestral form.
35	red-Hong-995	紅	ASR	D	赤 *chí* is the original word for 'red', and 紅 *hóng* is a later innovation, which is, however, reflected in almost all Chinese dialects, and is therefore almost impossible for the algorithm to be reconstructed.
36	rightside-YouBian-231	右邊	ASR	C	See above the entry for 'leftside-ZuoBian'.
37	small-Xi-997	細	ASR	H	The original word for 'small' is 小 *xiǎo*, so this is a clear independent innovation.
38	smell-Bi-1353	鼻	ASR	H	Most frequent word for 'to smell' is now 聞 *wén*. The form 鼻 *bí* is derived from the identical noun meaning 'nose'.

39	stand-Qi-528	企	ASR	H	企 *qǐ* has a specified meaning ‘to stand on the toes’, so we are dealing with parallel semantic shift here.
40	stone-ShiTou-968	石頭	?	C	Here, we find a typical problem for compound words in the history of Chinese. Given that all dialects reflect the form 石頭 *shítóu*, we need to ask ourself whether the compound form might not really have been the original form in the ancestor of all Chinese dialects. In the literature, the word occurs in Post-Hàn times (after 300 CE), which might coincide with the supposed separation of Chinese dialects.
41	suck-Shun-256	吮	GS	?	It seems that the word 吮 *shǔn* is regularly reflected in early sources of Hàn time (around 100 CE) in the meaning ‘to suck’, and it is thus much likely that this is a correct reconstruction.
42	sun-RiTou-669	日頭	ASR	C	日頭 *rìtóu* occurs in almost all dialect groups in this form. Similar to forms like 石頭 *shítóu* ‘stone’ or 月光 *yuèguāng* ‘moon’, we can ask ourself whether the late occurrence of these words in Chinese texts reflects linguistic reality.
43	tongue-SheTou-1143	舌頭	ASR	C	舌頭 *shétóu* is similar to the above-mentioned pervasive compound forms for ‘sun’, ‘moon’, ‘stone’, etc.
44	tooth-YaChi-871	牙齒	ASR	C	牙齒 *yáchǐ* is similar to the pervasive compound forms for ‘sun’, ‘moon’, etc., mentioned above.
45	who-ShaRen-305	啥人	ASR	H	啥人 *shárén*, literally ‘what person’ is extremely frequent in the Chinese dialects and has apparently been independently developed in different groups.
46	woman-NuGe-401	女個	ASR	C	女個 *nǚgè* is a rather rare form of the word for ‘woman’. Given the tendency of monosyllabicity, this form is also not likely to have been used to denote ‘woman’ in the ancestor of the Chinese dialects.
47	you-Ni-124	你	ASR	B	你 *nǐ* is not reflected in the archaic Mǐn dialects and generally an innovative form.