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

Gerhard Jäger
Institute of Linguistics
Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Johann-Mattis List
Department for Linguistic and
Cultural Evolution
Max Planck Institute for the
Science of Human History, Jena
mattis.list@shh.mpg.de

Current efforts in computational historical linguistics are predominantly concerned with phylogenetic inference. Methods for ancestral state reconstruction have only been applied sporadically. 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 outperforms 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.

ancestral state reconstruction – lexical reconstruction – computational historical linguistics – phylogenetic methods

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 (**?**), or family trees, using distance- (**??**) or character-based approaches (**????**), has entered the mainstream of historical linguistics. This is reflected in a multitude of publications and applications on different language families, from Ainu (**?**) and Australian (**?**) to Semitic (**?**) and Chinese (**?**). There is also a growing interest in the implications of phylogenetic analyses for historical linguistics, as can be seen from the heated debate about the dating of Indo-European (**????**), and the recent attempts to search for deep genetic signals in the languages of the world (**??**).

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 Fig. 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 (**?**), the direction of sound change processes (**?**), the original expression of concepts (**?**), or even linguistic and cultural aspects beyond the lexicon, such as ancestral color systems (**?**), numeral systems (**?**) or cultural patterns, e.g., matrilocality (**?**). 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. Our approach, in contrast, reflects the onomasiological perspective on the linguistic sign, as we try to infer the original *word* that expressed a given meaning. Since no commonly accepted name exists for this approach, 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 (**?**). For this purpose, scholars try to take known directional tendencies into account. These tendencies are usually based on the author's intuition, despite recent attempts to formalize and quantify the evidence (**?**). Following the classical distinction between *semasiology* and *onomasiology* in semantics, the former dealing with 'the meaning of individual linguistic expressions' (**?**: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.

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

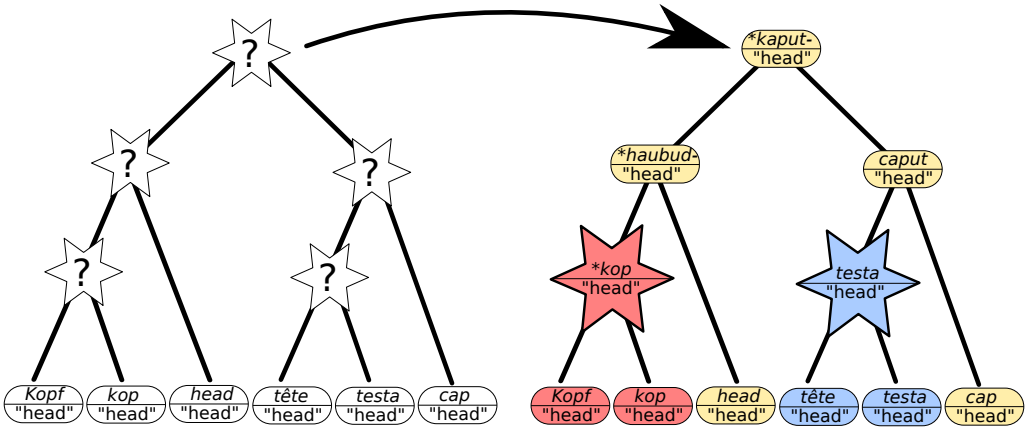

Figure 1
Ancestral state reconstruction: The graphic illustrates the key idea of ancestral state reconstruction. Given six words in genetically related languages, we inquire how these words evolved into their current shape. Having inferred a phylogeny of the languages as shown on the left of the figure, 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 possible 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.

This is partially reflected in the "topological principle in semantic [i.e. onomasiological, GJ and JML] reconstruction" proposed by **?**. 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; **?**), (2) determine a plausible phylogeny for the languages under investigation, and (3) use ancestral state reconstruction methods to determine which word forms were

[2] Notable exceptions include work by S. Starostin and colleagues, compare, for example, **?**.

most likely used to *express* the concepts in the ancestral languages in the tree. This approach yields an analysis as the one shown in Fig. 1.

Although we think that such an analysis has many advantages over the manual application of the topological principle in onomasiological reconstruction employed by **?**, 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 a 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 from the typical shortcomings of all data-driven disciplines, namely the shortcomings resulting from erroneous data assembly, especially erroneous cognate judgments, such as undetected borrowings (**?**) and inaccurate translations of the basic concepts (**?**) which are investigated in all approaches based on lexicostatistical data. The risk 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 for the inferences made in our tests, but also for the data we used for testing.

In the following, we 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; **?**, accessed on September 5, 2016),
2. Austronesian languages, as reflected in the *Austronesian Basic Vocabulary Database* (ABVD; **?**, accessed on December 2, 2015), and

3. Chinese dialect varieties, as reflected in the *Basic Words of Chinese Dialects* (BCD; **?**, provided in **?**).

All datasets are originally classical word lists as used in standard approaches to phylogenetic reconstruction: They contain a certain number of concepts which are translated into the target languages and 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 we used for our studies is available from the supplementary material.

The BCD database was used by **?** and is no longer accessible via its original URL, but it has been included in **?** and later revised in **?**. It comprises data on 200 basic concepts (a modified form of the concept list by **?**) translated into 23 Chinese dialect varieties. Additionally, **?** 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; 980 of these 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 under creolization than it does in normal language change). 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 **?**), each having a different ISO code which is also covered in the Glottolog database (**?**).

[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`.

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.

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 Fig. 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 Multitree (**?**, accessed on October 14, 2016). The age constraint for *Pali* is based on information from Encyclopaedia Britannica (2010, accessed on October 14, 2016). The constraints for *Old Gutnish* are taken from **?** and those for *Old Swedish* and *Old High German* from **?**. All other age constraints are derived from the Supplementary Information of **?**.
- **ABVD.** We only considered trees consistent with the Glottolog expert classification (**?**). This amounts to 213 topological constraints.

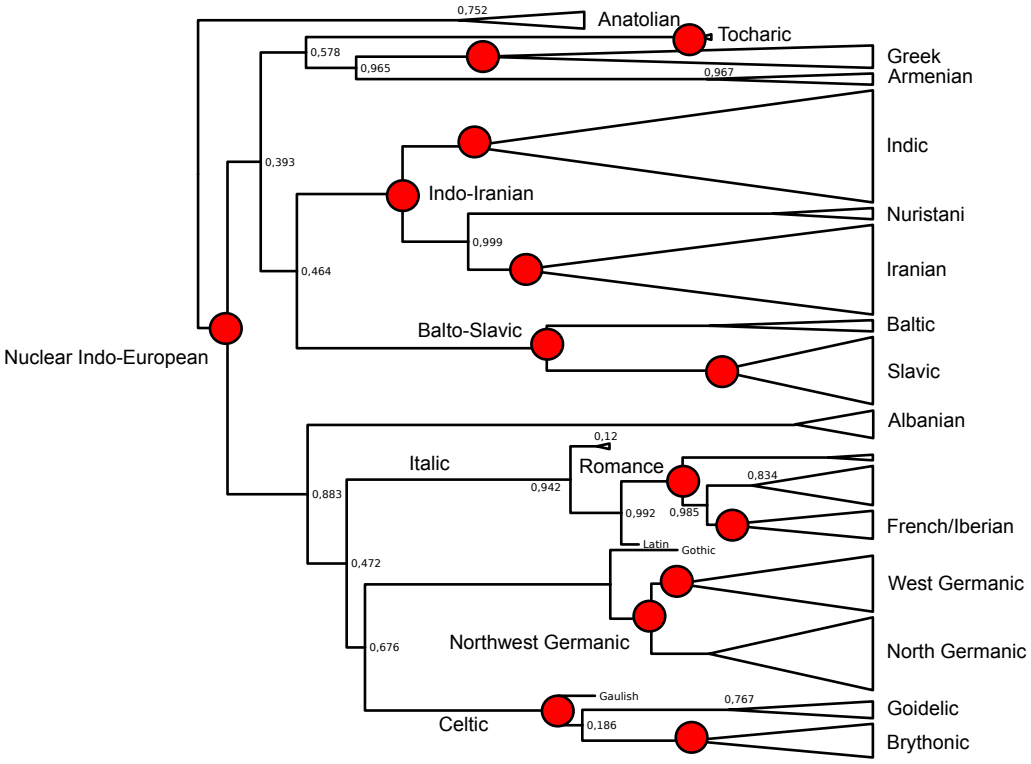

Figure 2
Maximum Clade Credibility tree for IELex (schematic). Topological constraints are indicated by red circles. Numbers at intermediate nodes indicate posterior probabilities (only shown if < 1).

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

Analyses were carried out using the MrBayes software (**?**). 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. **?**), with an exponential distribution with rate 200 as prior distribution for the variance of rate variation). Furthermore we assumed a birth-death model (**?**) 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*, retrieved on September 13, 2016; part of the software suite Beast, cf. **?**). 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 mutations 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 (**?**), (2) the minimal lateral network (MLN) approach (**?**) as a variant of Maximum Parsimony in which parsimony weights are selected with the help of the *vocabulary size criterion* (**??**), and (3) Maximum Likelihood (ML) reconstruction as implemented in the software *BayesTraits* (**?**). 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 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. The procedures are now presented for each algorithm in turn.

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 Fig. 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 leaves 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 leaf labels. In the example in Fig. 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 the sum of all penalties for all mutations in this scenario.[5]

[5] There is a variant of MP called *Dollo parsimony* (**??**) 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 presence and "0" absence of such a complex character, the weight of a mutation $1 \to 0$ should be infinitesimally small in comparison to the weight of $0 \to 1$. 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, emerge only once, recent empirical

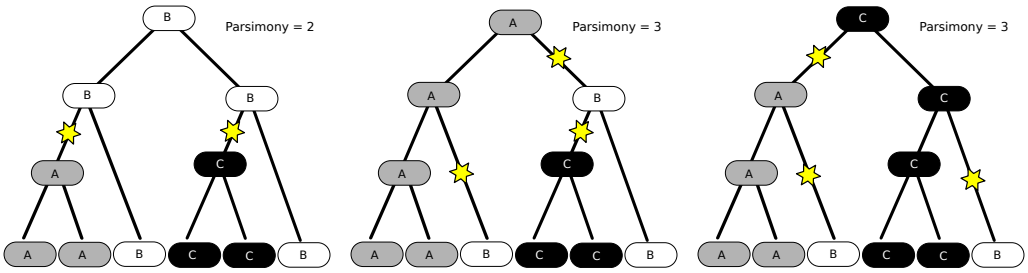

Figure 3
Complete character scenarios. Mutations are indicated by yellow stars.

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 leaf l of the tree is initialized with a vector $\mathrm{wp}(l)$ of length n, with $\mathrm{wp}(l)_i = 0$ if l's label is $states_i$, and ∞ else. (If l is polymorphic, all labels occuring at l have the score 0.)
- **Recursion.** Loop through the non-leaf 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 $\mathrm{wp}(mother)$ according to the rule

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

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

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**:

- For multistate characters, we used uniform weights, i.e., $w(i, i) = 0$ and $w(i, j) = 1$ iff $i \neq j$.

studies have uncovered that multiple mutations $0 \rightarrow 1$ can easily occur with cognate-class characters. A typical scenario is parallel semantic shifts. **?**, among others, 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 cognate according to IELex, 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. **?** 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 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 results 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 (**?**). In this form, it was also applied to linguistic data (**?**), and later substantially modified (**??**). 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 states 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 Fig. 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 primarily used to infer borrowing events in linguistic datasets, **?** 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 detail would go beyond the scope of this paper. For this reason, we refer the reader to the original publications introducing and explaining the algorithm, as well as the actual source code published along with the LingPy software package (**?**). To contrast MLN with 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

[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 (**???**), the best gain-loss ratio reported was 2:1.

[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 that 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 **?**.

the use of *binary characters states*: multi-state characters are not supported and can therefore not be included in our test.

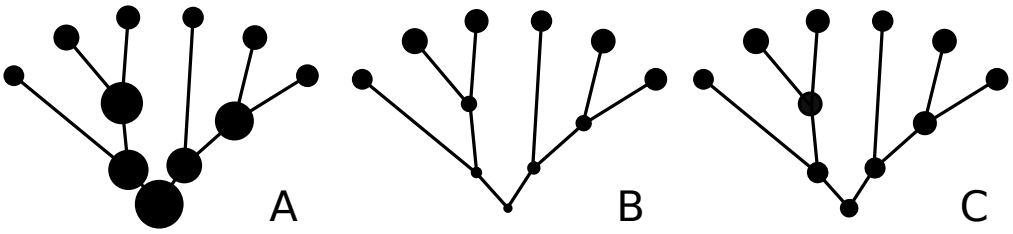

Figure 4
Vocabulary Size Distributions as a criterion for parameter selection in the MLN approach. A shows an analysis which proposes far too many words in the ancestral languages, B proposes far to few words, and C reflects an optimal scenario.

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 those on short branches if we assume that branch length 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 **?**, **?**, Section 15.7) for details. In the following 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 depends 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 loses it.

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]

$$\alpha = \frac{p_{01}}{p_{01} + p_{10}} \tag{2}$$

$$\beta = \frac{p_{10}}{p_{01} + p_{10}} \tag{3}$$

$$\lambda = -\log(1 - p_{01} - p_{10}) \tag{4}$$

$$P(t) = \begin{pmatrix} \beta + \alpha \cdot \exp(-\lambda t) & \alpha - \alpha \cdot \exp(-\lambda t) \\ \beta - \beta \cdot \exp(-\lambda t) & \alpha + \beta \cdot \exp(-\lambda t) \end{pmatrix} \tag{5}$$

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

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

a *continuous time Markov process*. This is illustrated in Fig. 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

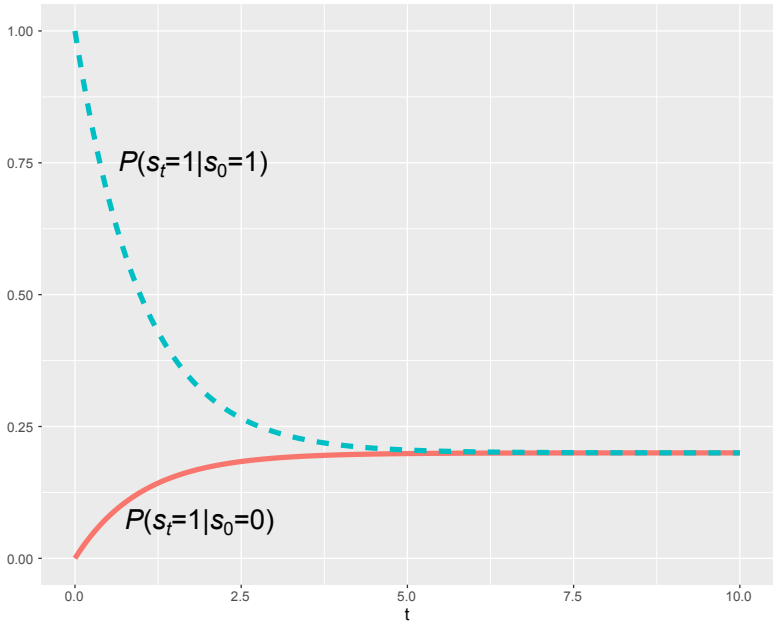

Figure 5
Gain and loss probabilities under a continuous-time Markov process.

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

$$\mathcal{L}(mother)_i = \prod_{d \in daughters} \sum_{1 \leq j \leq n} (P(t)_{i,j} \mathcal{L}(d)_j), \tag{6}$$

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 leaves. 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.)

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 assumed 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 datasets 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 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

[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.

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 or more 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 following we will deploy evaluation metrics established in machine learning to assess how well these two sets coincide:

$$precision \doteq \frac{|LCA \cap ASR|}{|ASR|} \tag{7}$$

$$recall \doteq \frac{|LCA \cap ASR|}{|LCA|} \tag{8}$$

$$F\text{-}score \doteq 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

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 **?**, 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 datasets 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 or 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 Fig. 6. The bars represent the average difference in F-score to the frequency baseline, averaged over all instances of the corresponding category across datasets.

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 choices introduced in the MLN algorithm or our version of Sankoff parsimony. Given that the MLN approach

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

was not primarily created for the purpose of ancestral state reconstruction, our findings do not necessarily invalidate the approach per se, yet they show that it might be worthwhile to further improve on its application to ancestral state reconstruction.

The impact of the other choices is less pronounced. Binary character coding provides slightly better results on average 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 penalties, are fixed by the researcher based on intuition and heuristics. ML, in contrast, uses branch length information, and it is based on an explicit probabilistic model of character evolution.

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

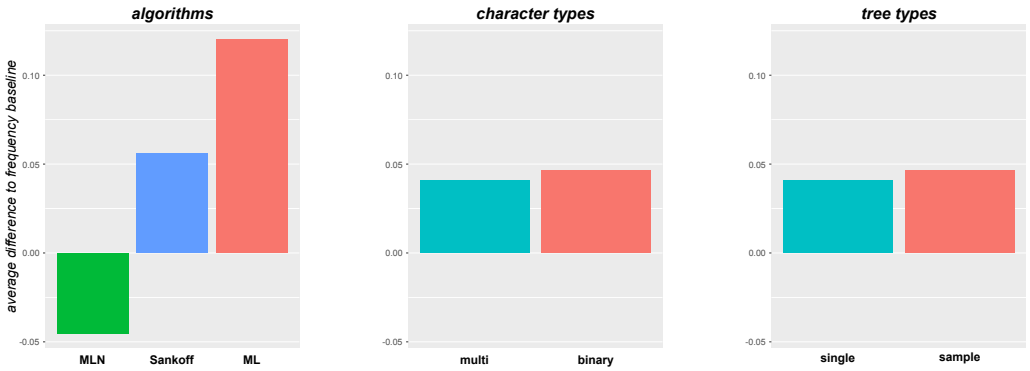

Figure 6
Average differences in F-score to frequency baseline

This point is illustrated in Fig. 7, which schematically displays ASR for the concept *eat* for the Chinese dialect data. The left panel visualizes Sankoff ASR and the right panel shows 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 Hakka and Yuè 粵 (also referred to as 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 surprisingly, both algorithms reconstruct *eat.Shi.1327* for the ancestor of the Southern dialects and *eat.Chi.243* for the 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

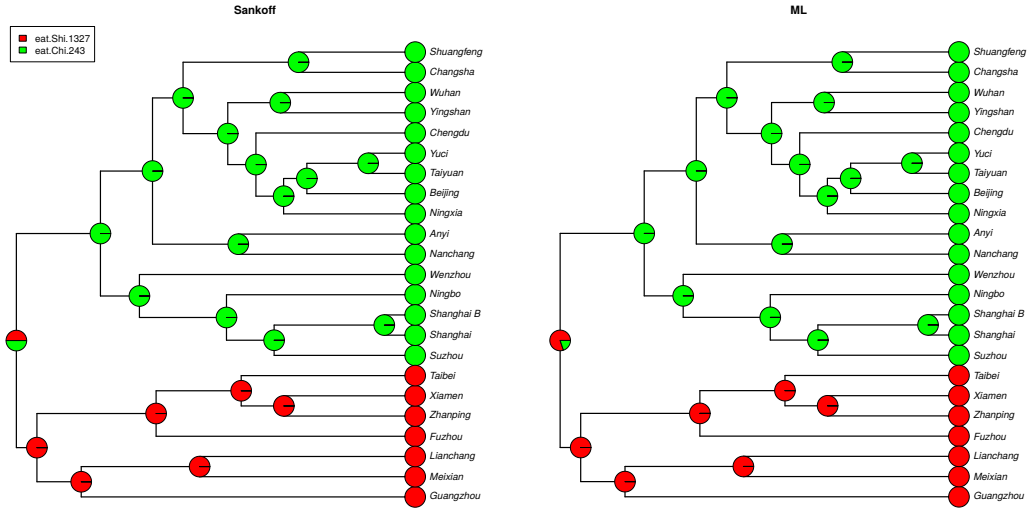

Figure 7
Maximum-Likelihood ASR and Sankoff Parsimony ASR for the concept *eat* for Chinese dialect data

ASR, on the other hand, 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 empirical, stochastic and data-driven approach to quantitative historical linguistics as compared to more heuristic and methods with few parameters.

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 specific difficulties and obstacles that the algorithms have to face when being used to analyze linguistic data. We therefore carried out detailed linguistic evaluations of the results proposed for IELex and BCD: we compared the results of the best methods for each of the datasets (Binary ML Sample for IELex, and Multi ML for BCD) with the respective gold standards, searching for potential reasons for the differences between automatic method and gold standard. The results are provided in Appendix B. In each of the two evaluations, we compared those forms which were reconstructed back to the root in the gold standard but missed by the algorithm, 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

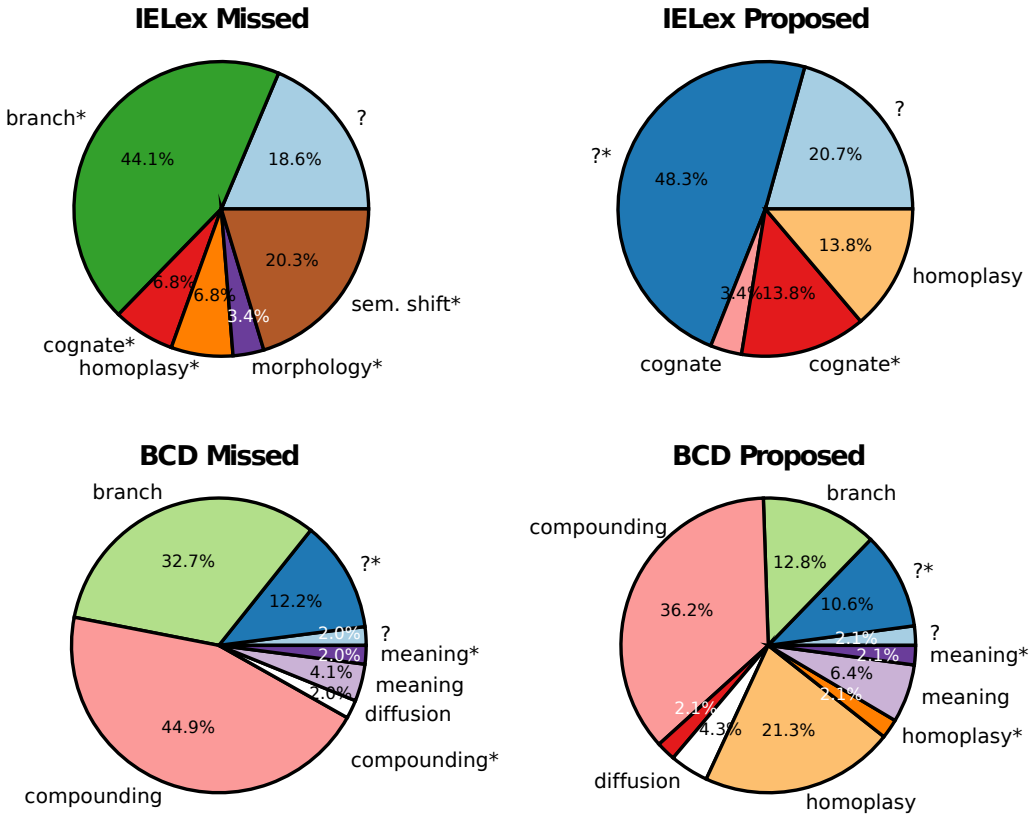

Figure 8
Detailed error analysis of the algorithmic performance on IELex and BCD. If a certain error class is followed by an asterisk, this means that we attribute the error to the gold standard rather than to the algorithm. For a detailed discussion of the different error classes mentioned in this context,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

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 differ slightly, and while a rather common error in IELex consisted in erroneous cognate judgments in the gold standard,[10] we find many problematic meanings that are rarely expressed overtly in Chinese dialects in BCD.[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 frequent problems in both datasets.

Fig. 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,

[10] See Appendix B1 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 B2 for details.

we can thus see that the majority of the 59 cognates missed by the algorithm can be attributed to cognate sets that are only reflected in one branch in the Indo-European languages and therefore do 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 (**?**:489). Second, the error statistics for Indo-European contain 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 **?**:86f confirms the reconstruction of the root, as it occurs in Proto-Germanic and Indo-Iranian, the meaning 'day' is by no means clear, as the PIE root *di̯eu̯- 'heavenly deity, day' is a more broadly reflected candidate for the 'day' in PIE (**?**:187f.).

Of the 29 cognates missed, 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 (**?**:93f) and generally considered a valid Indo-European root with the meaning 'worm, insect' (**?**: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, **?**), 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 that were interpreted as inheritance by the ML approach. As an example, consider the form *serp- (cognate class snake:E), which the algorithm proposes as a 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 **?**: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.

Turning to 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). However, here we cannot attribute these to questionable decisions in the gold standard, but rather to the fact that many Old Chinese words are often reflected only 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 and the first to branch 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 of compounds (口水 *kǒushuǐ* 'mouth' + 'water' = 'saliva'), it is only in the Yuè dialect Guǎngzhōu that it appears with 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*, in the latter, is given for Guǎngzhōu instead of *kǒu*. The differences in the data are difficult to explain, and we see two possible ways to account for them: (1) If *kǒu* was the regular term for 'mouth' in Guǎngzhōu in the data by **?**, 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, which all shifted the meaning of *huì*. (2) If *kǒu* is just a variant in Guǎngzhōu as it is 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. Multiple terms with similar reference potential were already present in the last common ancestor of the Chinese dialects. They were later individually resolved, yielding patterns that remind of *incomplete lineage sorting* in evolutionary biology (see **?** for a closer discussion of this analogy).

The problem of inherited variation becomes even more evident when we consider 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, heavily favored by the shift from a predominantly monosyllabic to a bisyllabic word structure in the history of Chinese (see **?** and replies to the article in the same volume for a more thorough discussion on potential reasons for this development). This development led 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 developed further in some dialect areas and yield a complex picture of patterns of lexical expression that are extremely difficult to resolve historically. Given that we find the words even in the most archaic dialects, but *not* in ancient texts of the late Hàn time and later (around 200 and 300 CE), the time when the supposed LCA of the majority 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 provided by Liú et al. (2007), with dialect surveys of the late 1950s, as given in Běijīng Dàxué (1964), we can observe how quickly Mandarin Chinese words have been diffusing recently: while we find only *rìtóu*[12] as a 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 between morphological processes like compounding and intimate language contact confronts us with challenging problems and may explain why 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 Fig. 7) in the Chinese dataset, where the branch lengths, which contribute

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

to the results of ML analyses, allow the algorithm to find the right answer. Another important point is the weakness of all automatic approaches and what we can learn from the 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 issues include first and foremost the problem of independent semantic shift, but also the effects of morphological change, especially in the Chinese data. **?** uses weighted parsimony with polarized (directional) transition penalties for multi-state characters 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 that 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* (**?**) 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 think it might be fruitful, both in classical and computational historical linguistics, to intensify the efforts towards semantic and onomasiological reconstruction.

Appendices

The appendices contain a list of all age constraints for Indo-European that were used in our phylogenetic reconstruction study (Appendix A) as well as a detailed, qualitative analysis of all differences between the automatic and the gold standard assessments in IElex (Appendix B1) and BCD (Appendix B2). They are submitted as part of our supplementary material.

Supplementary Material

All data used for this study, along with the code that we used and the results we produced, are available at `+++zenodo+++`.

If readers find error in the code or want to suggest improvements, they are cordially invited to file an issue via our GitHub repository at `+++github+++`.

The appendices A and B are submitted along with our supplementary material at GitHub and Zenodo.

Acknowledgments

This research was supported by the ERC Advanced Grant 324246 EVOLAEMP (GJ), the DFG-KFG 2237 *Words, Bones, Genes, Tools* (GJ), the DFG research fellowship grant 261553824 (JML) and the ERC Starting Grant 715618 CALC (JML). We thank our anonymous reviewers for helpful comments on earlier versions of this article, as well as all the colleagues who made their data and code publicly available.