Computational Phylogenetics

Claire Bowern,¹

¹Department of Linguistics, Yale University, New Haven, USA, 06520; email: claire.bowern@yale.edu

Keywords
historical linguistics, quantitative linguistics, evolution, phylogenetics

Abstract
Though there have long been links between research in historical linguistics and biological evolution, the last few years has seen growth in historical linguistic research which treats languages as evolutionary systems which can be investigated using tools from computational phylogenetics. Here, I explore some of the advantages and disadvantages of computational tools for historical linguistics. I describe the theory that underlies treating languages as evolutionary systems (in general terms), present the results of classifying languages lexically, and review some of the implications of the work.
1. INTRODUCTION

Linguistics, along with archaeology, anthropology, and genetics, is a crucial source of data for investigating about the past. Historical linguistics is a well developed field within linguistics (indeed, one of the oldest in linguistics), and many of the techniques that historical linguists use to investigate language change have been well studied. This includes work that looks at parallels between different types of evolutionary system, including those in language and biology. For example, it is often noted that parallels between linguistics and biology go back to Darwin (see comments in Atkinson & Gray 2005; Levinson & Gray 2012; Croft 2008). Perhaps most foundational in all this work is the concept of a “tree”, both as a convenient representation of relationships and a theoretically and empirically well motivated way to conceive of language split.

Computational (linguistic) phylogenetics is an important way to investigate both individual language histories and processes of change more generally. It is one of a suite of methods that historical linguists can use to find out about the past, and it has both advantages and disadvantages. In this review, I outline the theoretical underpinnings of phylogenetics — of “tree thinking” in contemporary computational historical linguistics — and describe the results of recent work building on this line of research.

First, it is worth considering what is meant by “computational phylogenetics”. There are several senses of the term as it is used in the literature. The first is tree inference: that is, the use of computational (particularly Bayesian) methods of determining internal branching of linguistic family trees. Some influential early work that uses phylogenetics in this way includes Atkinson et al. (2005); Gray & Jordan (2000); Holden (2002), and much recent work has followed. (Section 3.1 includes more details.) A second line of work (e.g. Dunn et al. 2011) studies the evolution of linguistic features and makes use of computationally-generated phylogenies, but does not infer the tree itself. Some of this work includes that discussed in Ladd et al. (2015). A third line of work uses trees or networks to cluster languages, but by similarity, rather than by strict historical descent. The latter might more accurately be considered a type of computational historical linguistics as broadly defined, and is not covered by the strict definition of ‘phylogenetics’ in, for example, Baum (2008). All three types of work are united, however, by a general approach to data gathering.
and analysis which stresses quantitative rather than solely qualitative arguments, explicit models of evolution, and specific hypothesis generation and evaluation.¹

In this review, I concentrate on the areas of computational phylogenetics which are particularly concerned with tree generation, though I do also briefly cover some of the more general ‘computational historical linguistics’ in Section 4. Croft’s (2008) Annual Review contribution on evolutionary linguistics provides some discussion of the theoretical basis of computational work in historical linguistics, as well as ties between evolution in biology and change in linguistics. Steiner et al. (2011) provides the most explicit recent framework for considering general computational approaches to historical linguistics.

2. Evolutionary thinking in historical linguistics

In this section, I concentrate on the theoretical underpinnings of phylogenetics. After all, any computational work on linguistic history must come to grips with whether methods violate assumptions about the data and processes that are being modelled. I begin with some background to linguistics as part of anthropology, particularly the emerging field of ‘evolutionary anthropology’ (cf Nunn 2011).

2.1. Language change as cultural transmission

Evolutionary anthropology is an approach to the study of the past that treats culture (including language) as an evolutionary system. That is, it models culture as collections of (more or less discrete) features which are shaped by history, the environment and ecology of their hosts (the humans that practise and transmit them), and other properties of the human mind and behaviour.² Some of the key questions of this approach to the past involve the relative roles of horizontal and vertical transmission — that is, how much influence language contact might have on how languages change over time. Another is the commonalities and differences between societies, the causes of change, and the ways that groups differentiate themselves from one another, and lineages split over time. These, and others, are very much the types of questions that have motivated historical linguists working in non-computational paradigms for a hundred years (see also Gray et al. 2013).

Croft (2008) provided an extensive review of these issues, particularly the role of Darwinian evolution, in which the focus of linguistic evolutionary studies was the transfer of work Dawkins and Hull on evolutionary theory into historical linguistics. This work particularly concerned the notion of a replicator, and the units in which linguistic information is transmitted. Since then, effort has been concentrated not so much in evolutionary theory, but rather in the application of analytical techniques from biology to specific problems in historical linguistics. That is, the theory has been led by the application of particular solutions to particular linguistic problems. It is therefore a good time to take stock of our current assumptions.

¹For a similar discussion of types of problems in evolutionary anthropology, see Jordan (2013).
²Mesoudi (2011); Richerson et al. (2001); Dediu (2011); Dediu & Cysoow (2013); Gray et al. (2010, 2007); Wedel (2006); Nunn (2011) provide overviews of this general research program, among others.
2.2. Criticisms of phylogenetics

An evolutionary approach to language is by no means universally accepted. Some recent critiques include Donohue & Denham (2011); Andersen (2006); Blench (2015), and Pereltsvaig & Lewis (2015). Other defenses of phylogenetics include Gray et al. (2007); Greenhill et al. (2010); Gray et al. (2013); Greenhill & Gray (2012); Jordan & Huber (2013); Bowern & Atkinson (2012), as well as of course, much work that simultaneously presents the problem and its theoretical approach.

Three types of criticism are commonly leveled at contemporary phylogenetic research; these are summarized here and further discussed in the following sections. The first relates to the degree of comparisons drawn between linguistics and biology. Critics of this view argue that because languages and species are different entities, there is no possibility of using the same methods to study both. One of the clearest critiques of this argument on the linguistic side comes from Andersen (2006); Blench (2015). Other work has borrowed metaphors from biology and cultural evolution but has explicitly rejected a general ‘evolutionary’ framework for language change (Blevins 2004). I take up this point further in Section 2.2.1 below.

A second type of criticism relates to the specifics of the application of phylogenetic methods to linguistic data. These criticisms usually point to the clear and long-standing issues raised in the literature on problems with lexicostatistics and glottochronology, such as how cognate items are identified, and how to distinguish shared innovations from shared retentions in modeling. I argue here that these criticisms are misguided; see further Section 2.2.2. That is, the critiques of lexicostatistics are valid, but computational phylogenetics does not share them. The third family of problems concern the type of data used, such as the heavy use of lexical data in constructing families rather than sound change or shared morphological innovations. These points are taken up in Section 2.2.4.

2.2.1. Biology and linguistics. Crucial to any ‘evolutionary’ enterprise in linguistics is whether the investigation is coherent for the domain under study. While academics in all fields have long turned to areas outside their immediate field of expertise as a way of seeking inspiration for solutions to problems within their own field, it has been a matter of debate whether the ‘inspiration’ of biology in historical linguistics involves treating one area simply as a metaphor for another. Perhaps the differences between domains are such that the transfer can never be more than metaphorical; or can we, in fact, extract meaningful generalizations from one domain and apply them directly to the other.

Crucial for the critics of the problem of whether biological methods are inappropriate for work in historical linguistics is whether the obvious differences between biological and linguistic systems are crucial for modeling. After all, cheesecakes and lasagnas are different foods, but we can still describe how to make both of them using recipes; they are both cooked in an oven, and both cut with knives after they are cooled. Their similarities at the level of chemistry outweigh the differences brought about by their ingredients. Similarly, a line of research in the history of science argues that differences in most aspects of evolutionary systems are not crucial for modeling (that is, do not invalidate modeling assumptions), as long as the following three criteria are met (cf. further Lewontin 1970; Sterelny 2006; Laland & Sterelny 2006; Wedel 2006).

1. Variation
2. Differential fitness [that is, selection]
3. Heritability
That is, systems must satisfy the three fundamental characteristics of a Darwinian system (Darwin 1859) in order to be studied with evolutionary models.

Systems must have variation in order to change. Systems with no variation are inherently stable. More specifically, variation is crucial because in an evolutionary system, the seeds of change are variation. Variation originates in random (spontaneous) mutation in individuals, and variants are then either neutral, or positively (or negatively) selected for. That is, variants Within linguistics there has been a lot of discussion of how to define the term ‘change’, including that change should be defined as the spontaneous creation of a novel variant in the mind of a single speaker (Hale 1998; Lightfoot 2002). But this focus on the creation of innovations belies the fact that the bulk of ‘change’ in linguistics comes from shifts in the relative frequency of variables, and not from innovation in a single speaker de novo.

Some work has looked at the idea of differential fitness of linguistic variables. This has been problematic, because it has relied on simplistic definitions of fitness. Some work has focused on definitions of fitness that primarily involve communication: for example, see the discussion in Kirby (1998). I see selection most usefully defined along two dimensions: a psychological/neurological dimension, which deals with perceptual, production, and processing biases in language; and a social dimension, which deals with the role of language as marking associations with particular social groups. That is, we might see variants propagating or not due to speaker internal factors, or speaker-external ones.

In sum, it is claimed, any system with these properties can be modeled in general terms using evolutionary methods, though the specifics of the models will vary. That is, the details of specific aspects of the model, such as the transmission mechanism and the relative amounts of horizontal and vertical transmission, are parameters of the model. Even within the domain of biology, there is a great deal of variation in the rates of change and the degree of horizontal versus vertical transmission. The ‘languages aren’t species’ trope applies equally well in biology: after all, gorillas aren’t orchids.³

Therefore, we need to be careful about the specifics of our models and what the assumptions are. We must be careful of what aspects of our research program crucially rely on a particular type of transmission mechanism, for example. Or that rely on specific properties of data. I doubt very much, for example, that Mendelian genetics will ever be particularly applicable to linguistics, Some models will fit our data better than others, and that of itself is likely to prove insightful.

Thus in summary, just because languages aren’t biological species, it doesn’t mean that they don’t have features that let us model their changes using evolutionary methods.

### 2.2.2. Problems with lexicostatistics and glottochronology

Early computational work in historical linguistics was concentrated in the fields of lexicostatistics and glottochronology. The former is figuring out language relatedness through linguistic similarities, while the latter is concerned with dating divergence through an assumption of a constant rate of change. Lexicostatistics is typically applied when comparative method analyses are not available. Consider, for example, Dyen, Kruskal, and Black’s (1992, 3) justification of Dyen’s lexicostatistical classification of Austronesian, quoted here at length:

³In taking this view, I argue against Andersen (2006) and Blench (2015), who take domain-specificity of models as crucial.
This validation is important because subgrouping the Austronesian languages by traditional methods is not feasible. Furthermore, even if it could be carried out, it is likely to yield only weakly supported results because the number of languages involved is large, many are scantily reported even when the reports are accurate, and very few have written records from earlier periods.

The problems with lexicostatistics and glottochronology are well described in the literature (see, amongst many others, Matisoff 1990; Campbell & Poser 2008; Crowley & Bowern 2010; Dunn 2015). The two most serious problems relate to the false assumptions that underlie their methods. The most relevant issues for us are the assumption of a single, constant rate of linguistic change on the one hand, and on the other, the use of linguistic distance as a proxy for linguistic relatedness. There are many observations that a single rate of change (approximately 18% of vocabulary replaced per thousand years) is simply a crude estimate.

Let us consider the method for lexicostatistics (Black 1997; Guy 2007, for recent treatments, compare). First, the linguist codes for superficial similarities between wordlist items (with no account taken of regularity or how similarities might arise). Here, the judgment of the linguist is paramount, and different linguists may come to different conclusions about thresholds for similarity. We then draw a tree by clustering the languages based on this similarity. Languages with higher percentages of shared items are grouped together. We then apply arbitrary (but motivated) cutoffs to define dialects, languages, subgroups, and families (Dyen et al. 1992). We then apply an exponential decay formula with a constant rate of change (given in 2.2.2) to estimate the amount of time since the languages diverged.

\[
(1) \quad t = \frac{\log C}{2 \log r}
\]

where \( t \) is the amount of time (in thousands of years) since the languages diverged, \( C \) is the percentage of shared features, and \( r \) is the retention rate (e.g., 0.805).

Now, there is no explicit theory of change behind this result. Implicitly, it, like all distance measures, is based on the premise that the more similar two languages are to one another, the more recently they diverged from a common ancestor. This is, of course, true to some extent, in that recently diverged dialects retain more features in common than members of different subgroups. But between those extremes, the method is highly sensitive to borrowing, chance similarities, and takes no account of regularity in correspondences.

A final reason why lexicostatistics (and, subsequently, phylogenetics) has been treated with suspicion is Dyen, Kruskal and Black's comment (quoted above) that lexicostatistics can be used where results from the comparative method will be unconvincing. Lexicostatistics is usually based on short wordlists, with superficial inspection for cognacy and little attempt to be consistent in the application of regular correspondences. Bayesian phylogenetic methods will not compensate for a lack of signal for relatedness.

Thus we should be careful about equating all types of computational work with each other. Lexicostatistics and glottochronology are deeply flawed: they are, in short, unsophis-
ticated treatment of complex data with clearly incorrect assumptions. But it does not follow that all computational approaches are equally flawed, especially computational approaches with different assumptions about language evolution.

2.2.3. Problems with interference from loanwords. Another claim (compare Donohue & Denham 2011; Pereltsvaig & Lewis 2015) is that phylogenetic methods are sensitive to loanwords, and because they usually do not exclude loanwords from consideration, the tree results are sensitive to them. However, a large advantage to Bayesian methods with lexical cognate data is that many data points go into the analysis. For example, Bowern & Atkinson (2012) included over 5000 binary characters for 194 languages. Because the individual input of any single character is very small, the models are robust to individual errors. Likewise, small numbers of undetected loans are unlikely to bias the results. They will make the signal appear more noisy and contradictory, however. Loans are thus treated rather differently by Bayesian phylogenetics than by human researchers applying the comparative method, who take sound patterns as indicative of relationship and then Marshall evidence of similar patterns elsewhere. Loans can lead to confirmation bias for human researchers, whereas for Bayesian approaches, extensive loans will decrease the levels of posterior support for subgroups in the tree. Further discussion of this, and demonstrations of how different treatment of loanwords in real data might affect results, is given in Bowern (forthcoming). The impact of loans on phylogenetic inference is a testable question. Greenhill et al. (2009) provides evidence form simulation studies that phylogenetic analyses are robust to loan levels up to about 30%. Bowern et al. (2011) has shown that average borrowing rates in basic vocabulary (of the type commonly used in phylogenetic studies) are low; of the order of 10% or less for most languages. Therefore, while the descent history of individual lexical items may include loans, the overall picture from Greenhill et al. (2009) and Bowern et al. (2011) is that loans may affect the classification of individual languages in some circumstances, but loan levels are very unlikely to lead to systematically incorrect classifications across a tree.

2.2.4. Problems with the type of data. A final set of criticisms are concerned with the input data used for computational analyses such as tree building. It is often claimed (Dunn et al. 2005, e.g.) that the ‘gold standard’ for language history is the Comparative Method, which works on establishing sound correspondences and from them inferring sequences of sound changes. Those sound changes are then used to construct trees of relationship, as well as to identify lexical items that might not fit the established correspondences (that is, to make arguments for morphological change and borrowing). Conversely, lexical data is assumed to be a poor indicator of genetic relationship, because words can easily be borrowed, and there are high-profile cases in the literature (not least the English language) where a substantial fraction of the lexicon has its source in borrowed material (Haspelmath & Tadmor 2009, cf). Whether it is licit to assume that a tree model appropriately captures lexical evolution is further discussed in Section 3.1 below.

Most phylogenetic work these days uses either lexical data (Bouckaert et al. 2012; Bowern & Atkinson 2012; Gray et al. 2009; Holden 2002) or grammatical features (e.g. Dunn et al. 2005). There is no reason to assume that lexical data changes in a fundamentally different way from other aspects of language: it’s still subject to variation, heritability, and selection. We see variation in word choice: for example, an English speaker can refer to the object in (2a) by any of the words in (2b), depending on register, dialect, use of
metaphor, irony, and so on. These words have been passed down in the history of English, not created *de novo* by the speaker, showing heritability and transmission. The words do not all have the same connotations, and would be appropriate in different circumstances, which demonstrates that there is some selection based on context.

(2) a. 

b. cat, kitty, feline, moggie, tabby, lion, mouser, kitten, ...

Historical linguists often suggest that evidence from shared innovations in sound change is the best material on which to base claims of subgrouping and relationship. I disagree. Sound change is useful in languages where there is extensive sound change, where the relative chronology can be inferred (allowing us to distinguish loans from inheritances and innovations), where there is sufficient information in the signal that we can tell genuine shared sound changes from parallel changes, and where the lexicon is sufficiently conservative that words stick around long enough to exhibit the sound changes that we want to study. Sound change data crucially also makes use of lexical data, and it presupposes the lexical relationships that are the explicit object of study for most Bayesian phylogenetic methods.5

Lexical data gives us extensive data with which to test relationship hypotheses. Each individual data point makes a small contribution to the analysis of the overall patterns, in contrast to the study of sound change, where — though the patterns are usually reinforced by the study of many lexical items — each branch is often supported by only a few changes. Moreover, sound changes are usually inferred simultaneously along with the tree. This is what gives us the information from the correspondences that a pattern reflects an innovation, rather than a retention.

Some work uses grammatical data. Longobardi, for example, has a number of papers using presence or absence of grammatical features to build trees of Indo-European. For example, Longobardi & Guardiano (2009, 1679) argues that “abstract syntactic properties are reliable indicators of phylogenetic relations.” And it has long been suggested that typological features might be a more reliable indicator of genetic relationships than lexicon, for remote language families.

Perhaps ironically, some of the reasons that Longobardi gives in favor of using parametric data are actually arguments against using it. For example, consider or core properties of evolutionary systems: variation, transmission, and selection. Longobardi & Guardiano (2009, 1686) defines parameters as being universally set within a language; that is, there are common to all speakers of a particular language. In that case, however, they do not vary,

---

5 Note that we need to distinguish the use of lexical data from the evolutionary model. For example, both Bowern & Atkinson (2012) and Wichmann et al. (2011) use lexical data to build trees. However, Bowern and Atkinson use 200 words of basic vocabulary, manually evaluated for cognacy by a linguist with 15 years’ experience in the relevant languages. ASJP uses Levenshtein distance based on substitutions for segments in 40 words, with no cognacy judgements. These are radically different ways of building trees, based on very different amounts of data, of very different quality.
and cannot undergo selection. Therefore either they do, in fact, vary, or their change should not be modeled by evolutionary models of the type considered here. A further problem is that parameters are dependent on one another, either because they are set dependently, in parallel, or if we take the proposal of (for example) Biberauer et al. (2014), hierarchically ordered. Both of these violate assumptions of independence among the characters that make up the input data.

Parametric (grammatical) data is highly subject to parallelisms. Grammatical data is usually coded as binary — that is, as a series of “yes/no” questions (for example, “Does the language have obligatory PRO-drop?”) Because there are only two values that such a feature can take, we expect languages that have no relationship to one another to share features. If the number of features used in comparison is small, or the features are highly dependent, the data will provide a poor signal for phylogenetic relatedness. Consider, for example, the small survey of questions in (3).

(3) Feature | English | German | Japanese | Bardi
---|---|---|---|---
Impersonal passives? | no | yes | yes | no
Dative case? | no | yes | yes | no
Verb final? | no | yes | yes | no

In this case, German and Japanese share grammatical features, as do English and Bardi, even though we know from lexical and other data that German and English are closely related, and not demonstrably related to either Japanese or Bardi.

3. Phylogenetic Methods

In this section I review some of the most typical methods for investigating topics in historical linguistics using phylogenetic principles. Note that I cover only truly phylogenetic methods: that is, methods that assume a model of evolution explicitly. I do not include extensive discussion of distance-based methods (where the similarity between pairs of languages is used as a proxy for how long ago they might have diverged). For methods relating to distance-based algorithms, including NeighborNets, see Bryant et al. (2005), and briefly Section 4.

3.1. Tree construction

The canonical research area for computational phylogenetics uses Bayesian computational methods to infer linguistic relationships. That is, we can use computational models to construct trees (or networks) from the linguistic data. Typically, this takes the form of assuming relatedness amongst the languages under study, and using lexical or grammatical data to investigate the internal relationships of the family.

A typical Bayesian tree construction procedure is as follows. Following the data gathering and coding (themselves not trivial tasks), a model is constructed. This model needs to be explicit about several different types of assumptions:

---

6For more information on these steps, see Dunn (2015).
The evolutionary model includes information about the assumptions about the behavior of characters and their transition from one state to another. For example, are the gain and loss of cognates equally likely? Is there a single rate of change across the tree, or are rates allowed to vary? The clock model describes the way in which the evolutionary model is used to infer the age of the tree. For example, do we use a strict clock, where the amount of change is wholly proportional to the amount of time that has elapsed? or is the presumed rate of change allowed to vary (and if so, under what conditions)? For example, a relaxed clock model might draw the clock from a distribution of rates. This is a more realistic assumption. We know that different languages have undergone different amounts of change since their split from their nearest neighbor, but we also know that there is both an upper and a lower bound of rates of change. No language has undergone no change, but languages do not change so fast that they become mutually incomprehensible to speakers within a speaker’s lifetime, for example. Bayesian models also require priors. These are assumptions which constrain the analysis in various ways. For example, we might have a prior on tree branching, which penalizes many very long or very short branches. Nodes might be constrained; for example, a group of languages could be constrained to be always grouped together, or bounds might be placed on the age of a node in the tree. Constraints of this type allow us to incorporate information which we already know from the real world in order to focus on the parts of the analysis that we wish to investigate. Finally, Bayesian models require setting up. The actual process of inferring the tree is done via a Monte Carlo Markov Chain. For a clear explanation of the process, see Dunn (2015); Dunn et al. (2005); Greenhill & Gray (2009).

It may seem problematic that we assume relatedness of the languages before applying the model. That is, it could be seen to be circular; after all, in working with the Comparative Method, the evidence for relatedness of the languages emerges from the process of comparison. However, unlike the comparative method, these methods separate assumptions of relatedness from the internal grouping. That is, arguments as to whether a group of languages are related use different types of evidence from arguments about how they are related to one another. Establishing relatedness will typically use shared, archaic characteristics (this is true in reconstructions using the Comparative Method as well). But internal subgrouping (again, just like in the comparative method) uses shared innovations. If we focus only on the most stable characters in the data, we don’t get a good sense of the evolution of items down the tree.

Secondly, Bayesian methods are sensitive to how treelike the data are. If applied on data which are non-treelike, most Bayesian methods will provide results that the data are a poor fit for a tree model (for example, the model might not converge, or the network-like structure of the data will be reflected in poor posterior probability values, or extremely long branch lengths between subgroups. For example, consider the tree reproduced in (5), Figure 4 of Bowern (2012). While lower level groups in this tree have posterior probability values close to 1 (reflecting a high degree of confidence in their fit for the data), the node joining the West and North groups has very low support (0.34). Results from Bayesian
analyses like this are indicative that a tree model is a poor fit for these data, either because
the processes that led to the languages diversifying were non-treelike, or (as is most likely
in this case) because the languages diverged sufficiently long ago that there is no longer
phylogenetic signal in the data beyond the lowest groupings.

3.2. Ancestral state reconstruction

Once the linguist has a tree of the languages under study, a variety of methods are available
to reconstruct features of those languages computationally. These methods are still at an
eyearly stage. In some ways, they use similar methods to inference to comparative method
type work: for example, majority rules, parsimony, and the like. Early work in ancestral
state reconstruction used arguments from parsimony to infer the value of states at interme-
diate nodes in the tree (Nunn 2011). For a typical argument from parsimony, consider the
trees in (6) below. The task is to infer the value of the trait at the root of the tree (labelled
\(X\)). If we assume that the trait value was 1 at the root, we need to assume a singel change
in the history of language \(B\). Conversely, if we assume that the root value was 0, we must
assume two, independent changes (one in the history of language \(A\), and the other on the
way two language \(C\).

A common method uses the ancestral state reconstruction algorithm of Pagel & Meade
(2004). Rather than directly estimate the values of characters are specific nodes in the
tree, Pagel and Meade estimate the \textit{transition rates} between states. These rates are then
mapped onto the tree, which allows us to probabilistically estimate the probability of a
character state value at a particular point in the tree. For example, a trait with two states
(e.g. presence or absence) has two transition rates which need to be inferred: \(q_{01}\) (the rate
at which a state becomes present) and \(q_{10}\) (the rate at which the trait is lost). Tracking
the estimations of the state values at points in the trees allows us to see the the confidence with which a trait is reconstructed to a particular node in the tree.

3.3. Correlating linguistic and non-linguistic features

A growing body of work uses language trees as a proxy for cultural descent groups. That is, some cultural anthropologists use language trees as a proxy for the vertical descent of cultural features, based on the assumption that most language transmission is intergenerational, rather than horizontal. They assume that since a language is a complex bundle of intertwined features, while individual pieces may be borrowed, it is unlikely that communities will shift their languages wholesale very often; thus a linguistic tree can be used as a proxy for a vertical descent history of the group, against which the effects of areal diffusion and horizontal transmission may be compared.7 Kushnik et al. (2014) study the evolution of land tenure norms across Austronesian using these methods. They compare the phylogenetic signal (that is, in broad terms, the degree of congruence between the traits and the tree they are presumed to have evolved on) to geographic signal. The phylogenetic signal is derived from mapping cultural traits to a linguistic tree, while the geographic signal is determined by the distance between the current locations of speakers of those languages.

Kushnik et al. (2014) use a linguistic phylogeny to model descent of non-linguistic features. Other work does not use the phylogeny, but rather includes language family as a proxy for vertical cultural descent group. Towner et al. (2012), for example, studies whether cultural traits predominantly evolve vertically or horizontally, by grouping languages into groups based on 1) language family and 2) geography, and comparing the goodness of fit of a model reliant only on geography, only on phylogeny, or some combination of the two. Their models are able to discriminate between cultural practices that are highly dependent on geography (such as the use of salt) versus those that are more associated with language family (and by association, vertical inheritance), such as burial positions. In other work, language affiliations are used to control for ‘Galton’s problem’ — that is, that sampling purely based on geography does not take shared history into account. Botero et al. (2014) provides another example in a worldwide sample (Towner et al. 2012 is focused on California).

4. Other work that uses quantitative methods

While thus far we have been concerned with work which deals directly with trees, it would be remiss not to discuss work that is closely related to phylogenetics, if not concerned with building or using trees directly.

4.1. Alignment and cognate identification

The examples of phylogenetics discussed above, including Gray et al. (2009); Holden (2002); Bowern & Atkinson (2012); Atkinson & Gray (2006), make use of cognate judgments man-

---

7Such work ignores both that there is a sizable quantity of horizontal transmission in language, (cf. amongst many others Thomason & Kaufman 1988), and that children acquire language from their peers as much as from their caregivers. That is, even though language may, on the whole, show treelike evolutionary features, it does not follow that the mapping between ethnolinguistic groups and languages is straightforward.
ually compiled by linguists. That is, the linguists examined lexical data to select cognate terms and build arguments for cognacy, according to principles of the comparative method. This manually curated data matrix then forms the input for comparative phylogenetic methods. This is a time-consuming process and is highly dependent on the expertise of a limited number of specialists, for most of the world.

Various work has tried to either remove the linguists’ cognate judgments from the equation, or at least to speed up the process. Rama et al. (2017), for example, is an unsupervised procedure for clustering words by cognacy. List et al. (2017) compares clustering (and subsequently alignment) procedures from several algorithms. Several algorithms match terms by the phonological similarity of the forms (or sometimes just the consonants). Under this method, forms that are above a certain (user-defined) threshold for similarity are deemed to be cognate. Other models make use of regularities in segment alignment within putatively cognate words to accept or reject cognate judgments. That is, words are deemed cognate if they exhibit regularities in sound correspondences. Bouchard-Côté et al. (2013) simultaneously learns lexical cognate judgements and sound correspondences, but requires a pre-existing tree. Hruschka et al. (2014) simultaneously infer segment alignments, sound changes, and the tree they imply. At this point, none of these models do as well as human coders, though Lists’ (2017) ‘InfoMap’ approach comes fairly close.

4.2. Inferring remote relationships

While the phylogenetic work described in Section 3.1 is concerned with family-internal classification, several papers have made use of phylogenetic methods to make arguments in favor of or against linguistic relationships at the level of the family. Here mention three pieces of work where the primary aim was inferring evidence for remote relationships: Dunn et al. (2005) on relationships among Papuan languages of New Guinea, Bowern (2012) on the languages of Tasmania (discussed and illustrated in (5) above), and Pagel et al. (2013) on the languages of Eurasia. Each article uses different methods to draw their conclusions.

Dunn et al. (2005, 2008) use trees created from grammatical data to examine the evidence for remote relationships among Papuan languages in island Melanesia. They argue that grammatical features are more stable than lexical ones and use a combination of phylogenies and correlations with geography to make the case that a number of Papuan languages are related beyond what can be shown by the comparative method. Pagel et al. (2013) uses reconstructed data from seven families of Eurasia to argue for a macro-family reaching from Indo-European to Eskimo-Aleut. They concentrate on correspondences among the words most likely to be indicative of remote relationships and suggest that a Eurasian superfamily began to break up approximately 14.5kya. Both are controversial, the former because long-term stability of structure features is unclear, and the latter because of problems with the family-level reconstructions on which the superfamily is based.

5. Conclusions

In conclusion, computational phylogenetics allows us insight into evolutionary processes in language. Phylogenetics has a firm theoretical base, both in biology and in other evolutionary systems, such as language. Cultural phylogenetics with language gives us new insight into human history over the last ten thousand years. While many of the criticisms of phylogenetics have been misplaced, such work is only as good as the data that it is based
on. Fancy maths will not save fairy tale cognate sets. Computational phylogenetics does not replace traditional linguistic work, such as the comparative method; rather, it gives linguists a powerful toolkit with which to investigate language history.

**SUMMARY POINTS**

1. Language and biology are both evolutionary systems, subject to parallel features
2. Trees are an appropriate model for language relationship in many cases.
3. Phylogenetic trees allow us to see where the model might not be appropriate.
4. Linguistic phylogenetics allows us to compare language histories with other types of population history.

**FUTURE ISSUES**

1. As more data for languages becomes available digitally, we should see improvements in the use of alignment algorithms; as these improve, this raises the possibility of inferring phylogenies directly from alignment data.
2. As phylogenetics in linguistics grows as a field, we should expect to see further development of language-specific evolutionary models. Just because language and biology can be modeled with the same conceptual toolset, it does not follow that linguists should not develop their own models within this frame. This should shed light on more of the ways in which linguistic and biological evolution differ.
3. Further phylogenetics trees for language families will allow us to better investigate the properties of trees, e.g. what causes language split, allowing us a better understanding of macro-level language change.

**DISCLOSURE STATEMENT**

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

**ACKNOWLEDGMENTS**

Work leading to this article was funded by the National Science Foundation grants BCS-0844550, BCS-1423711, and HSD-0902114. Thanks to Russell Gray, Simon Greenhill, and Quentin Atkinson for discussion of some of the issues raised here.

**LITERATURE CITED**


Bowern


Jordan FM. 2013. Comparative phylogenetic methods and the study of pattern and process in kinship. *Kinship systems: Change and reconstruction*:43–58


Wedel AB. 2006. Exemplar models, evolution and language change. The Linguistic Review 23