Borrowing, Character Weighting, and Preliminary Cluster Analysis in a Phylogenetic Analysis of the Ancient Greek Dialects*

Christina Michelle Skelton

Harvard Society of Fellows, Harvard University, Cambridge, MA, USA
cskelton@fas.harvard.edu

Abstract

Phylogenetic systematics is an increasingly popular tool in historical linguistics for reconstructing the evolutionary histories of groups of languages. One problem in applying phylogenetic methods to languages is that phylogenetic methods assume evolution takes place strictly by descent with modification, whereas borrowing between languages is common. This paper tests two different methods for addressing borrowing in phylogenetic analysis of language on a dataset representing the dialects of ancient Greek: character weighting and preliminary cluster analysis. Both methods show promise; they correctly recovered the subgrouping of the Greek dialects and were able to improve the resolution of the tree compared to the preliminary analysis. However, they recovered conflicting subgroupings of the West Greek dialects. This result is most likely due to a circular dialect continuum within West Greek. Using phylogenetic methods in situations which match their assumptions is crucial; for the West Greek dialects, phylogenetic network methods would be more appropriate.

* The author would like to thank the editor, Ron Kim, and the two anonymous reviewers, as well as Randy Diehl, Andrew Garrett, Scott Kominers, Richard Meier, Craig Melchert, Johanna Nichols, Don Ringe, Ed Stabler, and Brent Vine for their comments. Walter Gilbert first suggested the idea of cluster analysis. Matthew Bufano, Matthew Elkins, and Rebecca Hale assisted with coding. A large part of this research was conducted at the Center for Hellenic Studies, and the author would like to thank everyone, but particularly Greg Nagy, for their support and hospitality. This research was supported by a National Science Foundation Graduate Research Fellowship, grant number DGE-0707424 (2009–2011), a UCLA Graduate Research Mentorship (2012), and a UCLA Dissertation Year Fellowship (2013).

Keywords

phylogenetics – ancient Greek – Greek dialects

1 Introduction

Phylogenetic systematics, the set of methods originally developed in the biological sciences for reconstructing the evolutionary histories of groups of organisms, is an increasingly popular tool in historical linguistics for studying the development of language families and dialect groups (e.g. Nichols and Warnow 2008, Forster and Renfrew 2006). Phylogenetic methods can successfully be applied to the evolution of language families because both populations and languages evolve by descent with modification. Phylogenetics has several major advantages over traditional historical linguistic methodology, particularly in transparency and processing power. The data matrix that each phylogenetic analysis requires as input makes clear exactly what data the analysis was based on. The phylogenetic method chosen, be it an algorithm or an optimality criterion and search strategy, makes clear how the analysis arrived at an optimal tree or trees. Other analyses can be run after the fact to determine how well a given tree fits the data. Since the analysis is carried out by a computer, it is able to consider more data and run analyses faster than any human being ever could.

However, one major methodological problem in the application of phylogenetic systematics to language families is ‘lateral information transfer,’ or the exchange of information between different branches of the tree. Phylogenetic tree methods assume that evolution proceeds as a strictly bifurcating tree, with information transferred only from an ancestor to its descendants. However, language change though language contact is common—the linguistic equivalent of horizontal gene transfer.¹ There are a variety of means by which linguistic features can be transferred between unrelated languages, but in this paper, all types of lateral transfer of information will be referred to as ‘borrowing’ as a shorthand. Extensive borrowing may result in an unresolved phylogenetic tree or an incorrect tree topology.

Since a variety of linguistic situations can give rise to borrowing, a variety of different approaches are required to account for borrowing in phylogenetic analyses of language. Some linguistic features are more resistant to borrowing

¹ Horizontal gene transfer is exceedingly common in biological evolution as well; see, for instance, Keeling and Palmer (2008) and Zhaxybayeva and Doolittle (2011).

than others, a notion referred to as the 'cline of borrowability' (Sankoff 2002, Winford 2003, Matras and Sakel 2007). Morphology and syntax are the least susceptible to borrowing, while the lexicon is the most readily borrowed. Phonology is also relatively susceptible to borrowing (Sankoff 2002: 658). If it is the case that certain linguistic features were borrowed across different branches of the tree, then the results could be improved by giving these features less weight or removing them entirely.

Several studies have examined the impact of character weighting on phylogenetic reconstruction. Nakhleh et al. (2005) compared different phylogenetic reconstruction methods on an Indo-European dataset and examined in detail which phylogenetic characters were incompatible on the trees produced by the different methods. They concluded that, first, lexical characters were more easily borrowed than phonological and morphological characters, and second, that assigning appropriate character weights was important and had a significant effect on tree topology, though they did not reach any conclusions about what these weights should be. Barbançon et al. (forthcoming) tested the effect of character weighting on simulated data sets. Morphological characters were given a weight of 50 and lexical characters were given a weight of 1. When using Maximum Parsimony or Maximum Compatibility, they found that character weighting improved accuracy on data with relatively low homoplasy, but produced poor results on data with higher levels of homoplasy. Wichmann and Saunders (2007) examined the outcomes of using different phylogenetic methods to reconstruct the relationships among the Native American languages using typological data drawn from the World Atlas of Linguistic Structures. One method they tested was weighted Maximum Parsimony. Characters were weighted according to their rank, a rough measure of stability, with the lowest-ranked character receiving a weight of 1 and the highest-ranked character receiving a weight of 139. They found that weighting did not improve the overall tree topology, though it improved the bootstrap values for the correct nodes and weakened support for the incorrect nodes. However, the interpretation of these results is problematic, as Nichols and Warnow (2008: 807–808) note. In judging the outcome of the unweighted Maximum Parsimony analysis, Wichmann and Saunders (2007) base their conclusions on one of several optimal trees, presented in their Figure 4. However, the strict consensus of these trees, presented in their Figure 6, provides an acceptable tree topology. Therefore, the results and significance of the character weighting analysis are difficult to interpret.

However, lateral information transfer may also present a problem if languages or dialects are coded as discrete entities even when this may not be the most appropriate representation. As Barbançon et al. note:

There is a divide between the ‘between-species’ stochastic models of biological character evolution typically used in phylogenetic analysis, which usually assume monomorphism and also do not take population heterogeneity into consideration, and the ‘within-species’ models of population genetics, in which there is only partial geographical or reproductive separation between sub-populations, leading to polymorphism within sub-populations and the possibility that different samples of individuals from each of the sub-populations may exhibit varying evolutionary trees.

FORTHCOMING: 19

In other words, the phylogenetic analysis typically models the evolution of characters (for instance, DNA sequences or linguistic variants) among taxa (here, species or languages or dialects) in a way that assumes that there is no variation among the individual representatives of that taxon. Population genetics, on the other hand, assumes that different populations (species or smaller groups of organisms, or languages or dialects) may not be completely separate from each other, and that the individuals which make up the population may vary. This variation can cause different individuals from the same population to have different evolutionary trees.

Members of a dialect continuum which remain in close contact would present such a case. In this case, it may be appropriate to combine these languages or dialects and code them as a single entity. It is also worth noting that a group of languages that would be classified as a dialect continuum if one were to survey individual speakers may appear as distinct languages due to any number of reasons, for instance political boundaries, such as in the case of South Slavic; the development of literary languages; or poor attestation, which is likely the case for many Greek dialects. Such is the major pitfall of taxon selection in phylogenetic analysis of language: the need to impose granularity on a situation which is anything but discrete, or inappropriate granularity imposed by the vagaries of historical preservation.

Lastly, lateral information transfer may be present to such an extent that a strictly bifurcating tree is the wrong model for the data. In this case, it makes sense to analyze the data using a method other than phylogenetic tree models.

Thus, when performing a phylogenetic analysis, it is important to explore the data in order to come to an understanding of what linguistic situation could have given rise to borrowing. Any conclusions drawn from the analysis will be more secure, and the analysis may also uncover previously unknown developments in the history of the languages or dialects under study.

The present paper demonstrates this point using a data set where borrowing would be expected to be a problem: the dialects of ancient Greek. It first

presents a phylogenetic analysis of the Greek dialects, highlighting problems in the phylogenetic tree which may be due to borrowing. It then attempts to address these problems using two different approaches. The first approach uses different character weighting schemes to address the problem of whether certain borrowed linguistic features are distorting the data. The second approach uses a preliminary cluster analysis to determine how closely related certain taxa are in order to address the problem of whether the taxa have been properly defined. These two approaches both produce fully resolved trees, but with different tree topologies. A detailed look at the situation among the problem dialects reveals a heretofore unidentified dialect continuum. While both of these approaches give the appearance of having solved the borrowing problem—they both produce a fully resolved tree—only examining the data in detail gives an indication of which solution more accurately represents the complex distribution of dialect features in ancient Greek.

2 Background on the Greek Dialects

The Greek dialects are first attested with the Mycenaean dialect, used by the Bronze Age Mycenaean civilization across the southern Peloponnese, Crete, Boeotia, and Thessaly ca. 1400–1200 BCE (Shelmerdine 2008). However, the bulk of the Greek dialects are attested from the Archaic and Classical periods, ca. the eighth through the third centuries BCE, from the adaptation of the Greek alphabet to the spread of the *koiné* during the Hellenistic period. They are found across the Mediterranean, from southern Italy and Sicily to the north coast of Africa to Cyprus and the western coast of Anatolia. However, the geographic scope of this project is restricted to the Greek mainland, the Aegean, western Anatolia, and Cyprus, in order to best address the question of how the Greek dialects spread throughout these areas during the Bronze and Iron Age.

The Greek dialects are traditionally divided into four major dialect groups: Attic-Ionic, Arcado-Cypriot, Aeolic, and West Greek. These groupings are based on a variety of phonological, morphological, syntactic, and lexical features; for a thorough but succinct discussion, see Colvin (2007: 31–48). These dialect groups have a patchwork geographical distribution, and it was common for dialects to be in close contact with dialects of other dialect groups, as can be seen in Figure 1.² Attic-Ionic was spoken in a belt reaching from Attica

2 The base maps in Figure 1 and Figure 14 use data from Jarvis et al. (2008) and were developed in ArcGIS 10.3 for Desktop (ESRI 2014).

FIGURE 1 *Distribution of the Greek dialects in the first millennium BCE*

and the island of Euboea across the Cyclades to the western coast of Anatolia. Arcado-Cypriot was spoken in Arcadia, the central mountainous portion of the Peloponnese, and on the island of Cyprus. Mycenaean is commonly supposed to be most closely related to Arcado-Cypriot. Aeolic consists of Thessalian, spoken in Thessaly in northern Greece; Boeotian, spoken in Boeotia in central Greece; and Lesbian, spoken on the northwestern coast of Anatolia and nearby islands. Boeotian and Thessalian seem to have strong affinities with West Greek, while Lesbian seems to have strong affinities with Attic-Ionic. The debate as to whether Aeolic forms a true linguistic unity is still ongoing (e.g. Parker 2008). West Greek was spoken across northern Greece and the Peloponnese, the island of Crete, and other islands in the southern Aegean. Pamphylian, the dialect of Pamphylia in Asia Minor, does not fit neatly into any of these major dialect groups and may represent a mixed dialect (Colvin 2007, 47–48).

There are three major controversies in the development of the Greek dialects. The first is the first-order subgrouping of the major dialect groups. Prior to 1955, it was thought that the first major split divided Proto-West-Greek from Proto-East-Greek, which later developed into Aeolic, Attic-Ionic, and Arcado-Cypriot plus Mycenaean. However, Risch argued that the first major split in Proto-Greek divided Proto-North-Greek, which developed into West Greek and Aeolic, from Proto-South-Greek, which developed into Attic-Ionic and Arcado-Cypriot plus Mycenaean (Risch 1955).³

³ The current work generally follows Risch's view, though two alternatives are worth noting:

The second controversy is the geographical distribution of the dialects in the second millennium BCE. Mycenaean, which is mostly related to Arcado-Cypriot, was spoken in areas where other dialect groups are found in the first millennium BCE. West Greek dialects were spoken across the southern Peloponnese and on Crete, and Aeolic dialects were spoken in Boeotia and Thessaly. Arcado-Cypriot was only spoken in relatively remote and inaccessible places, the mountainous interior of the Peloponnese and the island of Cyprus. It is unclear when and how these Aeolic and West Greek dialects came to be spoken in the areas formerly occupied by Mycenaean.

The third controversy is the development of the West Greek dialects. One of the most significant events in the differentiation of the West Greek dialects was the development of the long vowel system after the first compensatory lengthening. The differences in question concern the existence of one or two series of long mid vowels, typically written in the Greek alphabet as $\epsilon\iota$ and $\omicron\upsilon$ for the higher set, and η and ω for the lower set. These long vowels come from the inherited Proto-Indo-European long vowels, a number of compensatory lengthenings, contractions, and the monophthongization of the diphthongs *ei* and *ou*.

According to Bartoněk (1972), there ultimately came to be three classes of long vowel systems: 'Mild Doric,' consisting of the Northwest Greek dialects and the dialects of the Saronic Gulf; 'Middle Doric,' consisting of West Argolic and Island Doric; and 'Severe Doric,' consisting of Cretan and Laconian. In the Mild Doric vowel system, there are four long mid vowels, in which $\epsilon\iota$ and $\omicron\upsilon$ represent the outcomes of the compensatory lengthening of *e* and *o*, contractions of *ee* and *oo*, and the monophthongization of the diphthongs *ei* and *ou*, and η and ω represent the long vowels inherited from Proto-Indo-European. Middle Doric has the same four long vowels, but η and ω represent the Proto-Indo-European long vowels as well as the outcome of the first compensatory lengthening, and $\epsilon\iota$ and $\omicron\upsilon$ represent the outcomes of the second and third compensatory lengthenings, the product of isovocalic contractions, and the monophthongization of the diphthongs *ei* and *ou*. In the Severe Doric vowel system, there are only two long mid vowels, written using η and ω (Ruijgh 2007: 394–395). Elean has its own distinct vowel system, which is not necessary to discuss in detail here.

Bartoněk believed that Severe Doric represented the original West Greek long vowel system. However, Ruijgh (2007) showed that the long vowel system

Brixhe et al. (2006), which rejects the idea that one can use 1st millennium BCE dialects to reconstruct 2nd millennium BCE subgroups, and Garrett 2006, who rejects the idea of Proto-Greek.

of Severe Doric must have arisen from the long vowel system of Mild or Middle Doric, since Severe Doric had of one set of long vowels, but Mild and Middle Doric had two sets of long vowels. However, many details in the development of the West Greek dialects are yet to be resolved (Colvin 2007: 44–45).

Thus, the standard view of the development of the Greek dialects suggests that there may have been two potential avenues for borrowing to affect the phylogenetic analysis. The first is for borrowing between adjacent but unrelated dialects, particularly between Lesbian and Ionic, and between Boeotian and West Greek. The second is for nontree-like evolution within the West Greek dialects, since there is no widely accepted subgrouping.

3 Phylogenetic Data Matrix

A phylogenetic data matrix consists of ‘taxa,’ which represent the entities under study, and ‘phylogenetic characters,’ which represent similarities and differences between the taxa. Each phylogenetic character has two or more ‘character states,’ or variants of the particular feature represented by the phylogenetic character. For this analysis, taxa represent dialects of ancient Greek. The phylogenetic characters represent phonological, morphological, lexical, and syntactic differences between the dialects. If a phylogenetic character represented the presence or absence of a given sound change, then that character would have two character states, representing whether or not that change had occurred. If a phylogenetic character represented the fact that different dialects used different words for a given meaning, then that character would have several character states, one representing each word. In total, the phylogenetic data matrix consists of 22 taxa, representing most dialects from the Greek mainland, the Aegean islands, Asia Minor, and Cyprus, and 85 phylogenetic characters. There are 40 phonological, 20 morphological, 3 syntactic, and 22 lexical characters, all function words.

The selection of taxa more or less follows Buck (1955: xi–xiii), with some important differences. As mentioned above, only dialects from the Greek mainland, the Aegean Sea, and the coast of Asia Minor have been included. Arcado-Cypriot is represented by Arcadian and Cypriot, with Mycenaean included as well. For Attic-Ionic, Ionic is represented by three separate taxa, representing West Ionic, Central Ionic, and East Ionic. These varieties of Ionic are separated by enough important differences, such as the outcome of the consonant clusters **ty*, **ky*, and **tw*, and the third compensatory lengthening, that is, the loss of *w* with compensatory lengthening of the preceding vowel in the sequence $V\{n, r, l, s, d\}w$, that it was necessary to code them as separate taxa. Another

taxon, of course, represents Attic. Aeolic is represented by four taxa, representing Lesbian, Thessalian, and Boeotian. Within Thessalian, the dialects of western and eastern Thessaly differ enough that they are worth dividing into two separate taxa, East and West Thessalian, representing the dialects of Pelasgiotis and Thessaliotis, respectively (Buck 1955: 150–151, Colvin 2007: 92). West Thessalian has a number of features in common with non-Aeolic dialects, most prominently West Greek, such as the genitive singular of o-stems in *-ou*, not *-oi*, and the present infinitive of thematic verbs in *-ein*, instead of *-emen*.

Within West Greek, Northwest Greek is represented by Elean, Locrian, and Phocian. The area around the Saronic Gulf is represented by Megarian, Corinthian, and West and East Argolic. However, Megarian was excluded from the phylogenetic analysis because it was almost identical to Corinthian, the only differences being a handful of character states which are known in Corinthian but unknown in Megarian. As with Ionic, there were enough important differences between West and East Argolic, such as the outcome of the second and third compensatory lengthenings, to necessitate splitting them into two separate taxa. The remainder of the Peloponnese is represented by Laconian. The islands are represented by Rhodian, Coan, Theran, and Cretan.

Pamphylian falls within the geographic area outlined above, so it was included in the data matrix for the sake of completeness. However, Pamphylian was excluded from the phylogenetic analysis. Several dialect groups, including Arcado-Cypriot, West Greek, and Aeolic may have contributed to its genesis, and it may also show influence from neighboring Anatolian languages, such as Lycian (Colvin 2007: 47–48). Given the circumstances, Pamphylian may represent dialect mixing. For this reason, Pamphylian is not the product of descent with modification, and would be inappropriate to include in a phylogenetic tree analysis.

It is common for a phylogenetic analysis to include one or more taxa to serve as an outgroup. An outgroup would ideally consist of the language or languages most closely related to Greek without being Greek, in order to determine which character states were ancestral and which were innovations. The outgroup also serves to determine where the tree should be rooted. Unfortunately, there is no ideal outgroup for ancient Greek. Phrygian, which is probably the language most closely related to Greek, is poorly understood. Armenian is the next most closely related language to Greek, but would also not serve well as an outgroup because its phonology is very complex, there is very little data for older forms of Armenian, and it is probably less closely related to Greek than previously thought (Clackson 1994). Vedic Sanskrit would be the next choice, but it is not as closely related to Greek as one would prefer, since innovations in Vedic Sanskrit may obscure ancestral forms. Macedonian may represent the

language most closely related to Greek, or it may represent another dialect of Greek (Hatzopoulos 2007, Méndez Dosuna 2012). In either case, the lack of data and the uncertain status of Macedonian makes it a poor candidate to include in a study which aims to improve phylogenetic methods. In light of these difficulties, no outgroup has been included. As a result, this phylogenetic analysis of the Greek dialects will not be able to determine which initial split divided the Greek dialects.

The phylogenetic characters are mainly drawn from Colvin (2007) and Buck (1955). When these sources were insufficient, data was drawn from Thumb (1909) and Bechtel (1921–1924) in general, and Dubois (1988), Egetmeyer (2010), and Brixhe (1976) for Arcadian, Cypriot, and Pamphylian, respectively.⁴ By primarily relying on two relatively recent and widely accepted handbooks, the data matrix aims to be uncontroversial, and phylogenetic characters have been excluded if their interpretation appeared too controversial. However, one of the great strengths of phylogenetic analysis is that it is easy to update the data matrix as new data and new research become available, and to test what effect other interpretations might have on the outcome of the analysis.⁵

4 Phylogenetic Methods

The phylogenetic method used in this paper is Maximum Parsimony. This method was chosen because in simulation studies of linguistic data, Maximum Parsimony appears to be the most accurate method (Barbançon et al. forthcoming). Maximum Parsimony has the additional advantage that it is possible to analyze the data afterwards to determine which phylogenetic characters support each branch of the tree. It would be worthwhile to test other types of tree models in future research, but that is unnecessary for the current work, because many of the problems and general conclusions will apply to all tree methods.

Maximum Parsimony works by assuming that the shortest strictly bifurcating tree, that is, the tree with only binary splits that implies the least amount of evolutionary change, is the one that is correct (Swofford et al. 1996: 415–416).

4 Even though Pamphylian has been excluded from the phylogenetic analysis, the taxon has been left in the data matrix in case it proves useful for others using this data set.

5 The author welcomes any such updates and inquiries. A detailed discussion of the phylogenetic characters included in the data matrix, along with the data matrix itself, can be found on the author's website at <http://www.christinaskelton.com/public/files/PhylogeneticCharacterDiscussion02.doc> and http://www.christinaskelton.com/public/files/Greek_Dialects_14_Mycenaean04.nex, respectively.

Conceptually, to accomplish this, the analysis first generates a large number of possible tree topologies, or possible ways to arrange the taxa into a tree given that each branch can only split into two daughter branches. The analysis then takes one such tree, and the phylogenetic data matrix that was the input. It uses the data matrix to determine where on the tree each of the phylogenetic characters in the data matrix must have changed in order to produce that arrangement of taxa. It then sums the total number of character state changes the tree required. This is the tree length. If the phylogenetic characters are weighted, it sums the weights of all the character state changes. At the end, the analysis selects the tree or trees which had the lowest tree length. The Maximum Parsimony analyses in this paper were carried out in PAUP* 4.0 for Windows (Swofford 1998) using the default settings.

5 Analysis

The remainder of this paper is devoted to testing the idea that borrowing may affect phylogenetic accuracy in different ways depending on the nature of the borrowing, and that different methods do a better or worse job of correcting for these problems. If the phylogenetic analysis of the Greek dialects is successful, one might expect it to recover the four major Greek dialect groups: Attic-Ionic, Arcado-Cypriot and Mycenaean, Aeolic, and West Greek. This would mean that the tree would be consistent with an initial split between Attic-Ionic and Arcado-Cypriot on the one hand, and Aeolic and West Greek on the other, even though the lack of an outgroup makes it impossible to root the tree. It should also fully resolve the subgrouping within any major dialect groups. However, the phylogenetic analysis may have problems with Boeotian grouping with West Greek and Lesbian grouping with Attic-Ionic. It may also not be able to resolve the subgrouping of the West Greek dialects.

The potential problems with Lesbian and Boeotian on the one hand, and the West Greek dialects on the other, reflect different underlying problems, and so should respond to different solutions. Lesbian and Boeotian developed as a part of one dialect group, then came into close contact with Attic-Ionic and West Greek, respectively, borrowing linguistic features heavily from them.⁶ Character weighting should be the best method of correcting for this type of borrowing, since it accounts for the fact that some features are more or less likely to be borrowed than others. However, the West Greek dialects may

6 García Ramón (1975), though Brixhe (2006) would disagree.

represent a dialect continuum. Thus, preliminary cluster analysis should be a better approach, since it can combine taxa which do not represent discrete entities. If the dialect continuum deviates too much from the model of treelike evolution, phylogenetic tree methods may no longer be appropriate.

One might question the validity of judging the results against an account which has increasingly come into question (e.g. Parker 2008), but it at least presents a convenient starting point which can be rejected if the phylogenetic analysis presents a compelling alternative.

6 Preliminary Phylogenetic Analysis of the Greek Dialects

This phylogenetic analysis, carried out using Maximum Parsimony with all characters given equal weight, produced seven optimal trees; see Figure 2. Optimal trees are defined as having the lowest possible branch length, so the reported statistics apply equally to all the trees. The results of the phylogenetic analysis closely match the results of traditional historical linguistic methods. The analysis identified the same four major dialect groups, and found that Attic-Ionic and Arcado-Cypriot, and West Greek and Aeolic, are more closely related to each other than the other possible arrangements. It also recovered the subgrouping of all the major dialect groups except for West Greek. Lesbian and Boeotian grouped with Aeolic, and not with Attic-Ionic or West Greek, respectively. Thus, only one of the potential borrowing problems affected the phylogenetic analysis of the Greek dialects: lack of resolution in the West Greek dialects.

7 Phylogenetic Character Weighting

Phylogenetic character weighting serves to improve phylogenetic accuracy by reducing the influence of characters which are more likely to be borrowed between branches of the tree. As mentioned above, phylogenetic character weighting works on the idea that there are structural constraints which make borrowing certain classes of features more or less likely, and it is already a well-established tool for improving phylogenetic accuracy in situations which involve borrowing.

This analysis tests the effectiveness of several different character weighting schemes at improving the outcome of the phylogenetic analysis of the Greek dialects. In the first set of tests, phonological, morphological, and lexical characters are weighted. Each case includes two runs, one with the characters of

FIGURE 2 Consensus tree for phylogenetic analysis of the Greek dialects with unweighted characters

that category given weight 2, and one with the characters of that category given weight 10. The other types of characters were given weight 1. The choice of 2 or 10 for weights is somewhat arbitrary, but weights higher than 10 produced results which were obviously wrong. In the second test, phylogenetic characters are weighted according to their consistency index (CI) in the outcome of the phylogenetic analysis in the preceding section. CI is measured on a scale of 0 to 1, where 1 indicates a character which changed the minimum possible number of times on the tree, and numbers less than 1 indicate characters which show some amount of homoplasy. The results are given in Figures 3–9.

The Greek dialect data does not support the assertion that weighting certain types of characters improves the results of the phylogenetic analysis. Every analysis except for the analysis with morphological characters weighted 10 (Figure 6) gave an incorrect tree topology according to the standard model of Greek dialect development, and that analysis was no better than the analysis with unweighted characters. Weighting phonological characters 2 (Figure 3) produced a tree in which Arcado-Cypriot and Aeolic were sister taxa. Weighting phonological characters 10 (Figure 4) produced a tree in which Arcado-Cypriot and Aeolic were sister taxa, and Attic-Ionic was the sister taxon to Island Doric. Weighting morphological characters 2 (Figure 5) and weighting lexical characters 2 (Figure 7) and 10 (Figure 8) gave trees where Lesbian was the sister taxon to Attic-Ionic, and where Aeolic was not a subgroup, but where Thessalian and Boeotian diverged individually from West Greek. This tree topology matches Parker (2008)'s model of the development of the Aeolic dialects, however. This suggests that supporting one model or the other simply depends on how much weight one gives to certain types of evidence. Weighting morphological characters 10 (Figure 6) did not change the tree topology or improve the resolution of the West Greek dialects compared to the run with unweighted characters, even though morphological characters are usually regarded as the most resistant to borrowing.

Prior studies of character weighting (Barbançon forthcoming, Nakhleh et al. 2005, Wichmann and Saunders 2007) have studied the development of language families and unrelated languages, where strong structural constraints against certain types of borrowing might be expected. However, the Greek dialect data may show that between dialects of the same language, these structural constraints are weaker or nonexistent, thus rendering different types of characters roughly equally likely to be borrowed. Barbançon et al. (forthcoming) found that data sets with moderate levels of homoplasy, which they define as 13 % of the lexical characters and 24 % of the morphological characters being homoplastic, character weighting resulted in poorer outcomes. 62 % of characters on the unweighted tree show homoplasy, so the results presented here

FIGURE 3 Consensus tree for phylogenetic analysis of the Greek dialects with phonological characters weighted 2

FIGURE 4 Consensus tree for phylogenetic analysis of the Greek dialects with phonological characters weighted 10

FIGURE 5 Consensus tree for phylogenetic analysis of the Greek dialects with morphological characters weighted 2

FIGURE 6 Consensus tree for phylogenetic analysis of the Greek dialects with morphological characters weighted 10

FIGURE 7 Consensus tree for phylogenetic analysis of the Greek dialects with lexical characters weighted 2

FIGURE 8 Consensus tree for phylogenetic analysis of the Greek dialects with lexical characters weighted 10

FIGURE 9 Tree for phylogenetic analysis of the Greek dialects with characters weighted according to CI

are perhaps unsurprising given the findings of Barbançon et al. (forthcoming). Interestingly, the different classes of characters each had comparable percentages of homoplastic characters. Based on the tree with unweighted characters, 64 % of phonological, 62.5 % of morphological, and 58 % of lexical characters were homoplastic.

Only the run with characters reweighted according to CI (Figure 9) improved the resolution of the tree while still producing a plausible tree, thus meeting the criteria for a successful reconstruction. This tree implies an initial split between Northwest Greek and Corinthian on the one hand, and the remaining West Greek dialects on the other. Within those dialects, there is then a split between Island Doric and the remaining West Greek dialects. Within that group, Laconian is the first dialect to diverge, then Cretan, and finally East and West Argolic. This treatment of the West Greek dialects will be discussed in more detail in the concluding section.

8 Preliminary Cluster Analysis

8.1 *Why Cluster Analysis?*

Preliminary cluster analysis serves to improve phylogenetic accuracy by identifying which taxa may not represent discrete entities from an evolutionary standpoint, and combining them into single taxa. If there appears to have been extensive borrowing between related taxa, it may be the case that the taxa are not separate entities which have undergone extensive borrowing, but two populations which never fully differentiated from each other and remained in contact. In this case, they should instead be represented as a single taxon in the phylogenetic analysis.

This new approach attempts to address this problem by performing a clustering analysis using Multidimensional Scaling on the single problematic branch of the Greek dialects, West Greek. The results will be used to condense a number of dialects into single taxa, and run the analysis again to see if the results have improved the resolution of the evolutionary tree.

From a phylogenetic standpoint, no phylogenetic analysis of language has yet developed a method to address borrowing at the level of taxa; character weighting addresses borrowing at the level of characters, and phylogenetic network methods address it at the level of phylogenetic methods. Yet there are clear grounds for rethinking the basis for dividing West Greek into dialects.

For West Greek in particular, an investigation of Cretan offers more insight into the line of reasoning which was used to define the individual dialects. Buck (1955: 171–172) lists the distinguishing linguistic features for Cretan, but notes that these linguistic features primarily describe the dialect spoken in

Gortyn, Knossos, Lyttus, Vaxus, and other areas of central Crete; the dialects of the eastern and western parts of the island are different. Buck discusses several differences, but ultimately rejects the idea that eastern and western Cretan represented different dialects. Bartoněk (1972: 91–92) begins with the assumption that Crete is a linguistic unity, but, in the course of his linguistic discussion, comes to the conclusion that there are enough differences between the different regions to treat them as separate dialects. Bile (1988: 10–12) notes that the conception of Cretan being divided into three parts geographically is a recent development, but that dividing the island based on physical geography and expecting the dialect geography to follow is misguided, and does not fit the linguistic situation on Crete. In all three cases, the starting point for defining the dialect was geography—the island of Crete constituted a single dialect, to be subdivided only if the linguistic data were strong enough.

If political divisions and physical geography have been taken as shortcuts for dialectal divisions, then there is room to consider the possibility that the traditional divisions of the dialects may not be wholly accurate from a linguistic point of view. Therefore, a more comprehensive examination of the linguistic data could provide reasonable grounds for splitting some dialects which may have been seen as geographic, political, or cultural unities, and for combining other dialects from areas which were geographically or politically distinct, but do not have sufficient linguistic differences.

8.2 *Cluster Analysis: Methods*

The analytical tool that can be used to test whether certain dialects should be grouped together as single dialects is cluster analysis. Cluster analysis is a general term for a large variety of methods, all of which attempt to group a set of entities based on some measure of overall similarity, so that the entities in a given cluster are more similar to each other than to the other entities. Cluster analysis is an extensive and diverse field; for an overview, see Osei-Bryson and Samoilenko (2014). The type of cluster analysis used here is Multidimensional Scaling (MDS). Multidimensional Scaling uses the distances between a set of entities to create an n -dimensional map showing the relative locations of the entities. In essence, Multidimensional Scaling attempts to solve the inverse problem of taking a map and being asked to measure the distances between a set of cities; Multidimensional Scaling takes the set of distances and attempts to recreate the map (Kruskal and Wish 1978: 7–8). Thus, a Multidimensional Scaling analysis of the West Greek dialects will produce a 2D image graphically representing the level of similarity or differences between the dialects in question. The investigator can then decide which level of similarity warrants lumping together two or more dialects as a single taxa.

FIGURE 10 *Scree diagram for the West Greek dialects*

Because Multidimensional Scaling produces a plot which shows overall similarities as distances, using MDS is neutral as to whether the dialects should be grouped together or remain separate. If all dialects in the input should be considered separate dialects, they should be spaced roughly equal distances from one another. However, if some dialects should be considered parts of the same dialect, they should appear much closer together, though the cutoff for which dialects are closely grouped enough to be combined is at the investigator's discretion.

The input to the MDS analysis consisted of the West Greek portion of the phylogenetic data matrix. It was not necessary to provide a distance matrix because the MDS analysis computes one for itself. The symbol "?" (missing data) in the phylogenetic data matrix was changed to NA (missing data) in the MDS analysis. Nonparametric MDS was used for the analysis because there is less risk of inappropriate assumptions about the relationship between proximities and distances affecting the stress values (Kruskal and Wish 1978: 76). The MDS analysis was implemented using the `isoMDS()` command of the MASS package in R.

For MDS, how well or poorly the results fit the data is measured through a value called 'stress.' Stress is essentially a measure of the badness of fit; the higher the number, the more poorly the analysis was able to fit the data. The stress values can also be used to determine the correct number of dimensions to use for the analysis. One way to test for the right number of dimensions is to create a scree diagram, which plots the number of dimensions against the resulting amount of stress (Holland 2008: 4). The point at which there stops being a significant improvement in the amount of stress is probably the correct number of dimensions. In this case, the correct number of dimensions is five.

FIGURE 11 *Multidimensional scaling analysis of the West Greek dialects*

The results strongly imply that the taxa have not been defined correctly. Instead, several sets of taxa should be grouped together as single taxa. These include Phocian and Locrian; East Argolic, West Argolic, and Laconian; and Theran, Coan, and Rhodian.

The groups of taxa identified above were then combined to produce single taxa. In general, when the dialects in a given group had different character states for a given phylogenetic character, the variant which was chosen was either the ancestral variant, or, if this was not clear, the majority variant.⁷

The Maximum Parsimony analysis of the data matrix with West Greek taxa combined gave a single phylogenetic tree. Thus, the clustering analysis was able to fully resolve the subgrouping of the West Greek dialects. The analysis suggests that the initial split in West Greek was between Northwest Greek plus Corinthian on the one hand, and Argolic-Laconian, Cretan, and Island Doric on the other hand. This would imply an initial split essentially between the dialects of the southern Peloponnese and the islands, and Corinthian and Northwest Greek. The treatment of the West Greek dialects will be covered in more detail in the next section.

⁷ Detailed notes on how the taxa were combined, as well as the data matrix with taxa combined, can be found on the author's website at <http://www.christinaskelton.com/public/files/WestGreekTaxaCombined01.doc> and <http://www.christinaskelton.com/public/files/GreekDialects14Mycenaeano2WestGreekCombined01.nex>, respectively.

FIGURE 12 Tree from phylogenetic analysis of the Greek dialects with West Greek dialects combined on the basis of MDS analysis

9 Phylogenetics, MDS, and the Prehistory of the West Greek Dialects

The purpose of these two sets of analyses, one comparing different types of character weighting, and one performing a cluster analysis prior to the phylogenetic analysis, was to determine which one of them, if any, offered the best way to resolve the lack of resolution produced by nontreelike data within the Greek dialects. The first analysis, which tested whether character weighting improves phylogenetic accuracy, found that there was no significant benefit to weighting phonological, morphological, or lexical characters, and, in fact, that weighting these classes of characters tended to produce incorrect tree topologies. Only reweighting characters based on CI resulted in a fully resolved tree. The second analysis, which used a preliminary cluster analysis, indicated that several West Greek taxa should be combined. This included Phocian and Locrian; East Argolic, West Argolic, and Laconian; and Theran, Coan, and Rhodian. When these taxa were combined and the phylogenetic analysis was run again, the result was a phylogenetic tree that was fully resolved. In short, both methods produced a fully resolved tree. However, resolution is the phylogenetic analogue of precision, and precision does not necessarily correlate with accuracy. In fact, these two fully resolved trees produced slightly different tree topologies, shown in Figure 13.

The two analyses only differ in the arrangement of three West Greek taxa: Argolic, Laconian, and Cretan. The analysis which reweighted characters according to CI shows Cretan as the sister taxon to Argolic, and Laconian as the sister taxon to this group. The analysis with preliminary cluster analysis combined Argolic and Laconian into a single taxon, with Cretan as its sister taxon.

These two scenarios are mutually contradictory, and it would appear that they cannot both be correct. Given the circumstances, it would be constructive to examine the data in more detail. Figure 14 presents a map of the dialects in question, using lines to show features shared between dialects. Judging by the thickest lines, which indicate the greatest number of shared features, the figure clearly indicates a dialect continuum which runs in a circular fashion around the Myrtoan Sea and the Sea of Crete, then branches off to Theran, Coan, and Rhodian in the southeast.

Thus, the analysis with characters reweighted according to CI and the analysis which used preliminary cluster analysis both obtained *a* right answer—they both recover tree topologies which are consistent with the circular dialect continuum—but not *the* right answer. The analysis with characters reweighted according to CI captured the portion of the dialect continuum running clockwise from Laconian to Cretan, since Argolic is most closely related to Cretan,

FIGURE 13 Comparison of West Greek dialects from phylogenetic analyses with characters reweighted according to CI and taxa combined based on MDS analysis

FIGURE 14 *Distribution of linguistic features in the West Greek dialects*

then to Laconian. The analysis which used preliminary cluster analysis captured the part of the dialect continuum running counterclockwise from East Argolic to Cretan, since Argolic and Laconian are most closely related to each other, then to Cretan, then to Island Doric. However, no phylogenetic tree analysis could recover a circular dialect continuum because it is incompatible with the model of a bifurcating tree.

It is worth searching for outside support for this new model of a circular dialect continuum among the West Greek dialects. In fact, such support exists. In discussing the origins of the Severe Doric long vowel system, Ruijgh (2007: 402) proposed that Severe Doric arose from either Mild or Middle Doric. The model of the circular dialect continuum predicts that the long vowel system of Severe Doric arose from Middle Doric, not Mild Doric. Phocian, Locrian, Megarian, Corinthian, and East Argolic, which show the Mild Doric long vowel system, form a neat geographical block in the north. However, West Argolic, which shows the Middle Doric long vowel system, is isolated from the other dialects which show the Middle Doric long vowel system, Theran, Coan, and Rhodian. The two dialects separating them, Laconian and Cretan, are both Severe Doric. Presuming that only dialects which were adjacent in the dialect continuum shared linguistic features, the simplest explanation is that West Argolic, Laconian, Cretan, and Island Doric were all originally Middle Doric, but that Laconian and Cretan later merged two sets of long vowels of Middle Doric into a single set. However, it should be noted that this is a simple and easily repeatable phonological innovation.

In fact, Ruijgh finds evidence that Severe Doric did indeed arise from Middle Doric. The earliest Cretan inscriptions show the long vowel system of Middle Doric, but around approximately 500 BCE, the vowel system changed to that of Severe Doric (2007: 403–406). Further evidence comes from the colonies Sparta is said to have founded, Thera and Tarentum. Thera, founded perhaps in the ninth century BCE, shows the Middle Doric long vowel system, while Tarentum, founded at the end of the eighth century BCE, shows the Severe Doric long vowel system. This implies that Laconian changed from the Middle to Severe Doric long vowel system perhaps around 800 BCE (406).

If the West Greek dialects do form a circular dialect continuum, it is also worth speculating on its origins and implications. Does its existence imply that West Greek came to this part of the Peloponnese only relatively recently, during the Iron Age, and so these dialects have not had the time to differentiate to the same extent as the other dialects? What aspects of the physical geography and trade routes allowed the circular dialect continuum to persist?

10 Conclusions

A number of conclusions about the development of ancient Greek and about linguistic applications of phylogenetic systematics can be drawn from the analyses presented in this paper.

First, the results from the different weighting schemes did not match what would be expected from the cline of borrowability proposed by various scholars. According to the cline of borrowability, weighting morphological characters should give the most accurate results since morphological characters are the most resistant to borrowing, followed by phonological and then lexical characters. However, the runs with phonological characters weighted 2 and 10 did not produce plausible tree topologies. All other weighting schemes produced plausible results, depending on which classification of the Greek dialects one adheres to. It will require further research to determine whether these results came about because dialects are mutually intelligible, and as a result have fewer structural constraints against different types of borrowing, or whether they are due to some historical idiosyncrasy in the development of the Greek dialects.

The results of the character weighting analysis did not provide a conclusive answer about the existence of an Aeolic subgroup, but they did show that the presence of the Aeolic subgroup depends heavily on the type of character weighting employed. Unweighted characters, characters reweighted according to CI, and morphological characters given weight 10 all recover the traditional

classification, which includes the Aeolic subgroup. Morphological characters given weight 2, and lexical characters given weight 2 and weight 10, do not recover the Aeolic subgroup, and instead group Boeotian and Thessalian with West Greek, and Lesbian with Attic-Ionic. Clearly, whether or not one accepts the existence of an Aeolic subgroup depends heavily upon how much weight one gives to certain types of evidence.

Preliminary MDS analysis appears to be a promising approach. It provided new insight into the development of the West Greek dialects and produced a fully resolved tree. It would be worthwhile to test this method further on additional data sets.

Two methodological conclusions can be drawn from the examination of the West Greek dialects. First, when attempting to improve a tree-based analysis, a better understanding of the data is more important than a better method. While reweighting characters according to CI and performing a preliminary cluster analysis did produce fully resolved trees, and even correct trees, these analyses missed an interesting and potentially critical basic fact about the development of the West Greek dialects.

Second, it is important to consider the question of what phylogenetic method, if any, is appropriate in a borrowing situation such as this one. Within phylogenetics, phylogenetic network methods which produce explicit networks and display borrowings as contact edges (branches connecting nodes which are not sisters) might be able to represent the circular dialect continuum within West Greek. Leaving phylogenetic methods behind, it may be useful to consider various methods drawn from dialect geography or population genetics.

In conclusion, in order to solve problems such as the development of the West Greek dialects, traditional historical linguistics and phylogenetic systematics must both expand beyond the model of a strictly bifurcating tree, in the sense of descent with modification leading to separation being the only paradigm used to model language evolution. It would be a victory for phylogenetic tree methods if they were able to definitively resolve the long-standing controversy of the subgrouping of the West Greek dialects. However, the fact that phylogenetic systematics and traditional methods encountered the same problems only serves to confirm that phylogenetic methods closely match the outcomes that can be obtained through traditional historical linguistic methods, while retaining advantages in their systematicity and their ability to handle far larger amounts of data than scholars using traditional methods. Clearly, the way forward is to test the effectiveness of phylogenetic network methods in analyzing borrowing situations such as the one found in the Greek dialects.

References

- Barbançon, François, Tandy Warnow, Steven. N. Evans, Donald A. Ringe, Jr, & Luay Nakhleh. Forthcoming. An experimental study comparing linguistic phylogenetic reconstruction methods. *Proceedings of a conference on language and genes, University of California, Santa Barbara, September 2006*. <http://www.cs.utexas.edu/users/tandy/nanterre-talk.ppt>.
- Bartoněk, Antonín. 1972. *Classification of the West Greek dialects at the time about 350 B.C.* Amsterdam: Adolf M. Hakkert.
- Bechtel, Friedrich. 1921–1924. *Die griechischen Dialekte*, 3 vols. Weidmann: Berlin.
- Bile, Monique. 1988. *Le dialecte crétois ancien: étude de la langue des inscriptions, recueil des inscriptions postérieures aux 10. (Etudes crétoises 27)*. Paris: Dépositaire, Libr. orientaliste P. Geuthner.
- Brixhe, Claude. 2006. Situation, spécificités et contraintes de la dialectologie grecque: à propos de quelques questions soulevées par la Grèce centrale. In Brixhe, Claude & Guy Vottéro, eds., *Peuplements et genèses dialectales dans la Grèce antique*. Nancy: Association pour la diffusion de la recherche sur l'antiquité, 39–69.
- Brixhe, Claude. 1976. *Le dialecte grec de Pamphylie*. Paris: A. Maisonneuve.
- Buck, Carl. D. 1955. *The Greek dialects*, 2nd edn. Chicago: The University of Chicago Press.
- Clackson, James. 1994. *The linguistic relationship between Armenian and Greek*. Oxford: Blackwell.
- Colvin, Stephen. 2007. *A historical Greek reader: Mycenaean to the Koiné*. Oxford: Oxford University Press.
- Dubois, Laurent. 1988. *Recherches sur le dialecte arcadien*, I–III. Louvain-la-Neuve: Peeters.
- Egetmeyer, Markus. 2010. *Le dialecte grec ancien de Chypre*. Tome I: *Grammaire*; Tome II: *Répertoire des inscriptions en syllabaire chypre-grec*. Berlin: De Gruyter.
- ESRI. 2014. *ArcGIS 10.3 for Desktop*. Redlands, CA: Environmental Systems Research Institute.
- Forster, Peter & Colin Renfrew, eds. 2006. *Phylogenetic methods and the prehistory of languages*. Cambridge: McDonald Institute Monographs.
- Garrett, Andrew. 2006. Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In Renfrew, Colin & Peter Forster, eds. *Phylogenetic methods and the prehistory of languages*. Cambridge, UK: McDonald Institute for Archaeological Research, 139–151.
- Hatzopoulos, Miltiades B. 2007. La position dialectale du macédonien à la lumière des découvertes épigraphiques récentes. In Hajnal, Ivo & Michael Meier-Brügger, eds., *Die altgriechischen Dialekte: Wesen und Werden: Akten des Kolloquiums Freie Universität Berlin 19–22 September 2001*. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck, 157–176.

- García Ramón, José Luis. 1975. *Les origines postmycéniennes du groupe dialectal éolien*. (Supplementos a Minos 6). Salamanca: Ediciones Universidad de Salamanca.
- Holland, S.M. 2008. *Non-metric Multidimensional Scaling (MDS)*. Unpublished ms., Department of Geology, University of Georgia, (<http://strata.uga.edu/software/pdf/mdsTutorial.pdf>).
- Jarvis, Andy, Hannes I. Reuter, Andrew Nelson & Edward Guevara. 2008. *Hole-filled SRTM for the globe, Version 4*, available from the CGIAR-CSI SRTM 90m Database (<http://srtm.csi.cgiar.org>).
- Keeling, Patrick J. and Jeffrey D. Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9(8), 605–618.
- Kruskal, Joseph B. & Myron Wish. 1978. *Multidimensional Scaling*. Thousand Oaks, CA: Sage.
- Matras, Yaron & Jeanette Sakel. 2007. *Grammatical borrowing in cross-linguistic perspective*. (Empirical approaches to language typology 38). New York: Mouton de Gruyter.
- Méndez Dosuna, Julián. 2012. Ancient Macedonian as a Greek dialect: a critical survey of recent work. In Giannakis, Georgios K., ed. *Αρχαία Μακεδονία: γλώσσα, ιστορία, πολιτισμός / Ancient Macedonia: Language, History, Culture / Macédoine antique: langue, histoire, culture / Antikes Makedonien: Sprache, Geschichte, Kultur*. Thessaloniki: Centre for the Greek Language.
- Nakhleh, Luay, Tandy Warnow, Donald A. Ringe, Jr., and Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an IE dataset. *Transactions of the Philological Society* 103. 171–192.
- Nichols, Johanna & Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2(5). 760–820.
- Osei-Bryson, Kweku-Muata & Sergey Samoilenko. 2014. Overview on cluster analysis. In Osei-Bryson, Kweku-Muata & Ojelanki Ngwenyama, eds., *Advances in research methods for information systems research*. New York: Springer, 127–138.
- Parker, Holt. 2008. The linguistic case for the Aiolian migration reconsidered. *Hesperia* 77. 431–464.
- Risch, Ernst. 1955. Die Gliederung der griechischen Dialekte in neuer Sicht. *Museum Helveticum* 12. 61–75.
- Ruijgh, Cornelius J. 2007. L'évolution des dialectes doriens jusqu'à la koina dorienn: le système des voyelles longues et la formation du futur. In Hajnal, Ivo & Michael Meier-Brügger, eds., *Die altgriechischen Dialekte. Wesen und Werden*. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck, 393–447.
- Sankoff, Gillian. 2002. Linguistic outcomes of language contact. In John K. Chambers, Peter Trudgill, & Natalie Schilling-Estes, eds. *The handbook of language variation and change*, Malden, MA: Blackwell, 638–668.
- Shelmerdine, Cynthia. 2008. Background, sources, and methods. In Shelmerdine, Cyn-

- thia, ed. *The Cambridge companion to the Aegean Bronze Age*, Cambridge: Cambridge University Press, 1–18.
- Swofford, David L. 1998. *PAUP**, *Phylogenetic analysis using parsimony (*and other methods)*, Version 4.0. Sunderland, MA: Sinauer Associates.
- Swofford, David. L., G.J. Olsen, P.J. Waddell, & David M. Hillis. 1996. Phylogenetic inference. In Hillis, David M., C. Moritz & Barbara K. Mable, eds., *Molecular systematics*, 2nd edn. Sunderland, MA: Sinauer Associates, 407–514.
- Thumb, Albert. 1909. *Handbuch der griechischen Dialekte*. Heidelberg: C. Winter.
- Wichmann, Søren & Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24. 373–404.
- Winford, Donald. 2003. *An introduction to contact linguistics*. Malden, MA: Wiley-Blackwell.
- Zhaxybayeva, Olga, and W. Ford Doolittle. 2001. Lateral gene transfer. *Current Biology* 21(7), R242–R246.