

Chapter 1

Paul Huff

April 13, 2009

A recent resurgence in interest has occurred in a sub-field of linguistics which has been relatively understudied over the last 50 years. Due to a variety of reasons, such as new statistical methods pioneered by scientists in other disciplines unrelated to linguistics and much faster and more capable computational resources than have ever been available to linguists in the past, lexicostatistics, the use of statistical and numerical methods to analyze language change and relatedness, has suddenly become much more compelling to a growing number of linguists.

As will be discussed in more detail below, mathematicians and biologists originally developed methods to understand the inter-relationships between species, and linguists have begun to use these methods to investigate questions of language relatedness and classification. Although language clearly doesn't operate on the same level as DNA in terms of its basic building blocks, these linguists have adapted measures of word and phoneme difference and similarity to the extent where they are able to then plug their results into the equations that biologists use for building family trees and therefore automatically build linguistic family trees. Interestingly, most of the equations used in practice to determine linguistic relatedness don't take advantage of most of the phonetic material available for such comparisons, or they rely on a

linguist to make such relatedness judgments by hand. As the key point of this thesis, I will below propose a method for using all the phonetic material available to make these comparisons as a way of improving the number of data points used in making these judgments. As a result this will make the comparisons more accurate. Before diving deeply into the subject, a review of the field of lexicostatistics and its current state will better situate the scope of the work presented here.

While over the last twenty years statistics have certainly been put to great use for a variety of other linguistic tasks, until very recently, they haven't been returned in earnest to what might be called the original questions that founded the field of linguistics itself: How did all known languages come to be as they are and how are they related to one another? While for roughly the first two hundred years of the field's life linguists were nearly universally interested in such questions, statistical methods have been anathema to the field for the last fifty years. The story of the return of statistical methods to the historical analysis of language is therefore as much of a story about the field of historical linguistics itself as it is a potential new way to shed more light on the subject.

Historical linguists have always used the historical-comparative method as the principal means to elucidate the relationships between languages. This method requires historical linguists to look at a group of languages' similarities and differences, construct a model of what their unattested ancestors must have looked like in order to better understand how languages have reached their current synchronic state, and then group the languages together into language families based on shared innovations from their ancestor languages (McMahon and McMahon 9–10). The method was developed for and used quite successfully to discover the relationships among the various Indo-European languages, and has been extended successfully to various other

language families throughout the world.

However, despite the historical-comparative method's successes, Calvert Watkins, the noted Indo-Europeanist, made clear that the method is not an end in and of itself, but instead is a means to the end goal of understanding the history of how the world's languages came to be as they are:

The reconstruction of Indo-European [via the historical-comparative method], the establishment, that is, of the grammar of that language to the best of our ability, is not our fundamental object, as it would be if we were writing a descriptive grammar of a known language. Rather, our ultimate aim is to write the linguistic history of known languages. We are seeking historical explanation for the grammar of languages accessible to us by observation or from written texts. Reconstruction is only a tool, a means to the end of understanding linguistic history.

Even if we were, by some miracle, handed a complete grammar of Common Indo-European as spoken somewhere in, say, 4000 B.C. (the date is meaningless), the work of the Indo-Europeanist would scarcely be done. In fact, it would be barely begun. For his task would be, then as before, to relate the facts vouchsafed him to the facts of attested languages: to construct hypotheses, and to demonstrate precisely how it is possible, within a linguistic tradition or traditions, for a language to pass from one system at one point in time to another system at a later point. (Watkins 101)

The historical-comparative method also has some problems, such as its inability to assign exact dates to language divergence events, and a lack of mathematical rigor to back up the assertions and hypotheses it asserts. Given that the goal of historical

linguistics is not historical comparison but, as Watkins describes, uncovering “the linguistic history of known languages,” and given some of the weaknesses of the historical comparative method, some historical linguists have tried using other, more mathematical, methods to investigate linguistic history (Watkins 101).

One of the most infamous mathematical methods used for discovering the relationships between languages is glottochronology, championed by Morris Swadesh and others. Swadesh sought to investigate the historical relationships between languages by creating lists for certain culturally universal lexical items. Glottochronology relies on the fact that so-called basic vocabulary items change less frequently than other vocabulary items, and so by finding basic, universal lexical items, the rate of change across various languages ought to be relatively constant. By looking at the number of items that were similar or cognate between the lists of basic vocabulary between two related languages and making the assumption of a constant rate of change of any given language, practitioners of glottochronology sought to establish with mathematical rigor a model for showing when languages descended from one another. The principal glottochronological equation is as follows (following McMahon and McMahon, 180):

$$t = \frac{\log c}{2 \log r}$$

where t = time depth in millennia, c = percentage of cognates and r = the glottochronological constant. Using this formula, Swadesh and others attempted to discern the exact split dates in millennia between languages and their ancestors (McMahon and McMahon 179–185).

Glottochronology’s methods were heavily criticized and eventually nearly universally rejected on a variety of grounds. Critics have attacked the methodology of list

construction, saying that there are very few basic lexical items that are universally present across cultures. They also attacked the rigor of the list construction process, since in some languages there are lexical-semantic pairs for which more than one word might fit. Since the accuracy of the lists themselves is suspect, the percentage of cognates portion of the glottochronological equation is suspect as well, as it depends on the lists themselves. Finally glottochronology's assumptions of a constant rate of change was most roundly criticized since using the formula leads to time depths which vary widely from known time depths. For language pairs like Tok Pisin and English the equation estimated the divergence as being much older than it actually was. For other language pairs, such as Old Norse and Icelandic and Old Armenian and Modern Armenian, the glottochronology equation predicts language deviation times which are vastly later than they actually were. These inconsistencies highlighted the difficulty of trying to model language deviation times with a single constant rate of change. Glottochronology's demise helped lead to a long disparagement of the use of numerical methods in historical linguistics (McMahon and McMahon 179–185).

Another relatively well known attempt at finding other ways of modeling language relatedness was proposed and refined by Joseph Greenberg. "Multilateral comparison" was the name he gave to the set of techniques and principles he developed in the 1950s and 1960s for creating genetic family trees of languages. Essentially he gathered lists of form-meaning pairs for sets of languages similar to Swadesh's. Then he used a complex set of mathematical formulas and factors for judging similarities across large numbers of languages at a time, rather than just single languages. At least ten different phenomena were considered as decisive, and were ranked in order. If a set of languages had a similar form-meaning pair, then the languages were able to be grouped together according to the mathematical weighting of the similarity

between the forms. More similarities between more pairs meant that there was much more likely to be a genetic grouping than if a group of languages simply shared one cognate between them (Croft, xiv-xx).

Although much of what Greenberg proposed seemed sound initially, his methods were also fairly universally criticized by the mainstream historical linguistic community, perhaps in part because the technology simply wasn't available at the time to allow him to perform his (quite rigorous) statistical calculations transparently, making the underlying processes seem somewhat opaque to his critics. Indeed, many critics criticized the sloppiness and seeming arbitrariness of his assumptions of similarity (Croft, xxvii-xxviii). They likewise criticized his use of error-ridden data, his lack of use of known genuine but dissimilar cognates, the large number of languages he investigated at a time, and several other methodological choices made by the methodology (Croft, xxvii).

Given the general and quite public rejection of both Greenberg's multilateral comparison and Swadesh's glottochronology by the historical linguistics community, for quite some time, no further mathematical models of language relatedness were seriously entertained. Given the rather dramatic and public failures of these kinds of mathematical models to gain traction among historical linguists one might ask why anyone would bother returning to mathematical methods for learning about linguistics. Calvert Watkins described the models the comparative method produces this way:

We must not forget, of course, that the reconstruction, the postulated grammar which is arbitrarily considered the initial point in the historical [comparative] linguistic process, is an artifact reflecting the contemporary state of intellectual development. As such it is subject to change, just as

all intellectual artifacts or scientific propositions are. . . . This mutability applies also to the model of kinship relations among a set of languages, the configuration of the family tree, which may also be modified—like any scientific proposition—by new data. (Watkins 100)

In the previous Watkins quote and in this one, while averring the legitimacy of the reconstructions created by the historical comparative method as valid scientific models, Watkins is clearly stating that the constructions of the historical comparative method are simply a tool in the historical linguist's tool belt (though admittedly the predominant one for the last 200 years). Additionally, Watkins clearly states that new scientific insight could certainly lead to a revision of the model of language relationships produced by the army of linguists that has been toiling away with the historical comparative method for centuries.

In recent years there has been a rekindling of interest in the use of statistical methods in the pursuit of historical linguistics to provide just such an added mathematical perspective as Watkins hinted might be useful to the model of language relationships already established by the historical comparative method. While linguists attempted to distance the field from the use of bad statistical methods such as glottochronology, biologists and mathematicians have in the mean time developed statistical and computational methods that are quite useful for answering specific questions about how species are related to one another and the specifics of those relationships (including the modeling of the timing of splits between species and the modeling of their family tree relationships). These methods and the questions they answer have clear parallels in linguistics, and recently a number of linguists have begun investigating them, creating a lexicostatistical renaissance in the field. By so doing, linguists are augmenting the already well-established history with the new data that Watkins said might help

	Language 1	Language 2	Language 3
Language 1	x	5	2
Language 2		x	3
Language 3			x

Table 1: A trivial example distance matrix showing the cognancy judgments between 3 languages as might be used by McMahon and McMahon or the ASJP project. The similarity between Language 3 and Language 1 is 2. Only half of the matrix needs to be computed, since the two halves are mirror images of one another. This saves computation time. After the matrix is generated, various statistical methods are used to compute trees and networks based on the information in the matrix. The difference between the McMahon and McMahon and ASJP methods and the method I propose is in the generation of the number which represents the similarity of two languages.

revise the configuration of the established history of languages.

Among others, a distinct program of lexicostatistical research which has proven quite fruitful is being developed by a few groups independently of one another. Two of these groups, McMahon and McMahon and the Automated Similarity Judgment Project (hereafter ASJP) have separately pursued a nearly identical approach. Both projects essentially use the same computational methodology to investigate the inter-relationships between languages. First, they obtain pair-wise distance measurements between a group of languages. Having generated pair-wise similarity measurements for every pair of languages in the group in question, they generate a distance matrix of all the languages in question (see table 1). They are then able to use off-the-shelf bioinformatic software to walk this distance matrix they have created and generate tree structures (McMahon and McMahon 89–118, Brown et al.). The key difference between the ASJP project’s research and the program of research undertaken by McMahon and McMahon appears to be in the method for determining the pair-wise distance between the languages.

In order to compute the distance between two languages, McMahon and McMa-

hon use a database constructed by Dyen, Kruskal and Black which is a large set of Swadesh lists that have been generated for the Indo-European languages (McMahon and McMahon 96). For each pair of languages in the database, a distance number between the pair of languages is generated based on the percentage of words in the two lists that are cognates. The decision of whether any two lexical items are cognate is pre-determined by hand based on the judgments of historical linguists and encoded in the database (McMahon and McMahon 97–98). McMahon and McMahon then further refine the items in question by using a set of criteria for determining which lexical items are less likely to be borrowed through contact and create two sublists, one which is highly likely to be affected by borrowing and therefore less stable over the long run, and one which is more likely to be unaffected by borrowing and therefore more stable over the long run. They then recompute distance matrices based only on those sublists from the Dyen, Kruskal and Black database which are stable and use the resulting tree structures to refine their automatically generated trees (McMahon and McMahon 107-109).

The ASJP project takes a different approach on the subject of distance matrix generation. After having collected Swadesh lists in nearly 2000 languages from all over the world, they devised a procedure which allows them to create trees for these languages using an automated process to generate the distance matrices (Holman et al. 3). For each set of two languages in the whole pool of available languages, the individual items of the Swadesh list are compared using the following process: two words with the same meaning are judged as similar if two consecutive consonants are identical between the two words, even if there are vowels between them (Holman et al. 3). The percentage of words which are similar between a language pair is used as a starting point for calculating the language pair's distance value. Borrowing and

accidentally similar phoneme inventories are then controlled for by subtracting the percentage of words which don't have the same meaning but are judged as similar by the program from the original percentage (Holman et al. 3–4). The final number is subtracted from 100% and with this percentage, the ASJP project obtains a distance for each pair of languages in their database (Holman et al. 4).

Once they have obtained distance matrices for a set of languages both McMahon and McMahon and the ASJP project then use similar techniques from computational biology to generate phylogenetic trees. McMahon and McMahon try a variety of tree- and network-building methods (McMahon and McMahon 89–175) including neighbor-joining, but the ASJP project uses only neighbor-joining to examine the trees generated by their distance matrices. Neighbor-joining works by selecting pairs of languages with the smallest distance first as a root node and then successively repeating the act of finding the smallest distance between languages to add branches to the tree (Paradis 100). There are several off-the-shelf software packages which perform neighbor-joining on data-agnostic distance matrices, and both groups make extensive use of them. McMahon and McMahon further use network building software to construct tree-like structures from the distance matrices that take into consideration places where it might be less statistically easy to build a dichotomous tree branch (McMahon and McMahon 139–175).

One place where both of these models somewhat naively treat linguistic distance is between individual lexical items. Once again, the McMahon and McMahon method uses a binary one or zero to model whether two languages have cognate form-meaning pairs for a particular slot, the cognate decision having been made before hand by historical linguists (McMahon and McMahon 96-98). The ASJP method uses a binary metric based on whether or not the first two consonants of a form-meaning pair are

identical between languages. An easy place to improve on both methods seems to be to take into consideration all the phonetic material available between two form-meaning pairs and compute a distance metric which is continuous and not just a binary “yes or no” number. In the case of McMahon and McMahon’s methodology, this will remove the necessity for a human cognancy judgment, making the number be based on known information (the phonetic strings of each language), rather than a potentially subjective cognancy decision. In the case of the ASJP’s methodology, using a distance metric based on all of the phonetic material in a string rather than just the first two consonants will allow a richer model based on more data points.

I propose using such an automatically generated distance metric to determine how closely related two phonetic strings are for any given semantic slot shared between languages. The distance between languages is then to be derived from the phonetic distance between individual strings in the languages in question. Given that phonetic distance metrics are crucial to this process, a review of some phonetic distance metrics which have been used for similar purposes is helpful.

Several different phonetic distance metrics have been proposed in the literature. Brett Kessler gives a good overview of some of these metrics in the context of multilateral comparison (5–6). The first metric, C_1 -place, derived by Kessler and Lehtonen, looks at the place of articulation feature for the first consonant of the two words being compared and, based on how far apart the places of articulation are for the two consonants, assigns a value to the phonetic distance of the two words. The second metric Kessler describes is P_1 -Dolg, created by Baxter and Manaster Ramer, based on work by Dolgopolsky. Dolgopolsky divided phonemes into classes based on two features: place of articulation and manner of articulation. The P_1 -Dolg distance metric looks at the first phoneme of the two words in comparison and if they are from

the same class of phonemes considers them equivalent, and otherwise considers them non-equivalent (Kessler 5).

A sort of cross-mixing of the two methods was investigated by Kessler in two more metrics, C_1 -Dolg and P_1 -place. Both of these metrics operate similarly to their above described counterparts; however, C_1 -Dolg uses the Dolgopolsky consonant classes to determine distance between the first consonant of each of the two words in question, instead of just the place of articulation feature, and P_1 -place uses the place of articulation feature with the first phoneme of each word in question, regardless of whether it's a vowel or consonant. P_1 -voice uses just the voice feature of the first phoneme to assign a distance score to two words from different languages, giving them a distance of 0 if the voice feature of the first phonemes are the same and 1 if they're different. (Kessler 5).

Kessler's C^* -DolgSeq attempts to exploit more material than just the first consonant of a word to generate a distance between two words. By aligning the consonants of the two words in question and then adding to the distance score of the two words if two aligned consonants aren't from the same Dolgopolsky class, a sort of composite distance score is obtained. Lastly, Kessler describes C^* -DolgCross, an attempt at fixing some of the problems that arise in C^* -DolgSeq: since aligning consonants between two words where one of the words is shorter than the other ends up with potential alignment problems (for example, in the case of epenthesis, where an extra consonant is inserted into the word, or in the case of *albero* and *arbol* in Italian and Spanish, which have the same vowels which have been reordered), the C^* -DolgCross distance score is computed by comparing the Dolgoposky class between all possible pairs of consonants between the two words, and averaging them together (Kessler 5-6). None of these distance metrics outperforms any of the others in the context

of the multilateral comparison tests in which Kessler them (Kessler 8). This should probably be unsurprising, given that none of the metrics uses more than two phonetic features, most of them use only one phoneme and that the solutions to problems arising to alignment issues that arise in the multi-phoneme metrics are relatively naive. None of these seem to be a competitive distance metric with the two proposed by McMahon and McMahon and the ASJP project.

Grzegorz Kondrak gives a survey of a different set of algorithms for determining the phonetic distance between two words which take into account many of the problems encountered by the metrics which Kessler describes (Kondrak 2–3). The first algorithm for the calculation of distance between two words that Kondrak describes was proposed by Covington. Covington’s algorithm groups phonemes into three classes: consonants, vowels and glides. Covington then assigns a distance to each of the possible differences in classes between two phonemes (see table 2). Additionally, deletions and insertions of phonemes between the two words being compared are also assigned penalty scores which are added to the overall distance measurement for the two words (Kondrak 2).

Kondrak also briefly describes two other distance metrics given by Gildea and Jurafsky, and Nerborne and Heeringa respectively. Rather than just using a single feature or a phoneme class, both of these distance metrics use binary feature vectors to represent the words being compared. A distance metric called the Hamming distance is used in both of these metrics to penalize a substitution from one phoneme to another, while inserts and deletions of phonemes are penalized somewhat arbitrarily by both (Kondrak 2).

In contrast to these metrics, Kondrak proposes a new algorithm for determining the similarity between two phonetic strings, called ALINE. ALINE has been shown to

Difference	Penalty
Identical consonants or glides	0
Identical vowels	5
Vowel length difference only	10
Non-identical vowels	30
Non-identical consonants	60
No similarity	100

Table 2: The penalties assigned by Covington’s distance calculation to varying classes of differences (following Kondrak 3). While the actual feature values aren’t taken into consideration, the method does make sensible judgments about which differences between phonemes should be seen as being worse. Covington himself described the method as being preliminary.

perform well under a variety of circumstances (Kondrak 60-64). ALINE works over multivalued feature vectors representing the words in question; that is, where Gildea and Jurafsky’s and Nerbonne and Heeringa’s methods both use feature values that are either on or off, Kondrak’s ALINE has a varying fraction between 0 and 1 for the multiple values of each of the features considered. Additionally, ALINE, rather than computing the distance between two words, computes the words’ similarities, giving higher scores for words which are more similar and lower scores for words which are less similar (Kondrak 3).

ALINE computes the similarity of two words by examining all the possible combinations of phoneme pairings between the two words, and then giving larger positive scores to phonemes which are similar, and large negative scores to phonemes which are dissimilar, and smaller negative scores to cases where an insert or delete has clearly occurred. By creating a matrix of similar segments, an optimal alignment can then be obtained from the segments in question, using a dynamic programming algorithm (Kondrak 3–4). Kondrak’s algorithm has been shown to work well under a variety of language types, and of all the algorithms discussed so far is the only one

that takes into consideration all of the phonetic material encapsulated in a word using multi-valued features (Kondrak 60-64).

Since Kondrak’s algorithm for phonetic similarity clearly seems the most robust and most accurate, it seems to be an excellent candidate for slotting in the step of determining the percentage of cognates shared between two languages which both McMahon and McMahon and the ASJP project do. In the case of McMahon and McMahon, using ALINE replaces the work done by linguists, eliminating some subjectivity in the results, and in the case of ASJP, the algorithm is improvement over the “first two consonants” heuristic for cognancy that is currently in use, since the similarity metric takes into consideration all the phonetic material in the lexical-semantic pairs, instead of just two phonemes from each.

As the ASJP project does, I control for accidental similarity between the phonological inventory of two languages by simply computing the phonological similarity between all of the lexical-semantic pairs which don’t match (i.e. compare the phonetic similarity between the word for *dog* in one language with the word for *milk* in the other) and dividing the total similarity score for the matching lexical-semantic pairs by the accidental similarity score. Once I have a matrix of similarity scores between pairs of languages I can then use off-the-shelf biological methods to examine the matrix, as do McMahon and McMahon and the ASJP project, and construct genetic trees to classify the languages as related or not. The tree-like signal of the classifications can be compared with known family trees for Indo-European languages and with the trees generated with other automatic methods, such as ASJP, to see how well the generated trees work as they should.

Using ALINE in this way, I can further refine the automated lexical statistical models of the past to hopefully add new, useful information to understanding of the

relationship between languages, which, as Calvert Watkins said, should help us to better “write the linguistic history of known languages” and give new data to help refine the models of the past.

References

- Cecil H. Brown, Eric W. Holman, Søren Wichmann, and Viveka Velupillai. Automated classification of the world's languages: A description of the method and preliminary results. *STUF Language Typology and Universals*, 61(4):285–308, 2008.
- Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, , and Dik Bakker. Explorations in automated lexicostatistics. *Folia Linguistica*, 42(2):(in press), 2008.
- Brett Kessler. Word similarity metrics and multilateral comparison. *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 6–14, 2007.
- Grzegorz Kondrak. A new algorithm for the alignment of phonetic sequences. *Proceedings of the First Meeting of North American Chapter of the Association for Computational Linguistics*, pages 288–295, 2000.
- Grzegorz Kondrak. *Algorithms for Language Reconstruction*. PhD thesis, University of Toronto, Toronto, Canada, 2002.
- April M. S. McMahon and Robert McMahon. *Language classification by numbers*. Oxford University Press, Oxford, 2005.
- Emmanuel Paradis. *Analysis of Phylogenetics and Evolution with R*. Springer, New York, 2006.
- Calvert Watkins. Language and its history. *Daedalus*, 102(3):99–111, 1973.