Computational and Quantitative Approaches to Historical Phonology

Brett Kessler

The Oxford Handbook of Historical Phonology
Edited by Patrick Honeybone and Joseph Salmons

Print Publication Date: Nov 2015
Subject: Linguistics, Historical Linguistics, Computational Linguistics
Online Publication Date: May 2015 DOI: 10.1093/oxfordhb/9780199232819.013.030

Abstract and Keywords

This chapter surveys methods of analysing phonological change that rely on computers because they require lengthy operations, mathematical precision, and reproducibility. Applications include techniques for discovering and verifying sound correspondences, modelling the course of sound change, computing the most likely genetic tree consistent with a set of innovations, testing the significance of the phonetic evidence for genetic relationship between languages, and exploring the relationships between dialects via quantification of phonetic and phonological differences.

Keywords: computational phonology, quantitative phonology, dialectometry, cladistics, statistical significance, genetic relatedness, modelling

9.1 Introduction

COMPUTATIONAL historical phonology is a difficult field to delineate. In fact, it can be argued that there is no such discipline. There are no core methodologies that are taught in every linguistics department and routinely employed by all historical linguists; indeed, there might not be any computational methods that the majority of historical linguists even consider valid. Difficult as it is to define the state of the art in computational historical phonology, this chapter seeks to provide a general overview of threads of computer-assisted research that have been pursued by several researchers for several years. It focuses primarily on computerized methods that use phonological criteria such as sound correspondences and sound similarity to investigate genetic relations between languages. A secondary focus is on phonetic comparisons between language varieties without regard to their genetic relationships.

A glaring omission is any discussion of the computer simulation of linguistic change, because Wedel (this volume) is dedicated to that important field, which would be almost inconceivable without computer models. For some other areas of research for which computers are often vital, see Yu (this volume) on experimental methods and Maguire (this volume) on corpus phonology.

(p. 134) 9.2 Comparative Method

The comparative method as described in textbooks sounds straightforwardly algorithmic. The following recipe is a synthesis of several descriptions; see for example Fox (this volume) for more detail, and the papers in Durie & Ross (1996).

- Data collection: the linguist collects words that express the same or related concepts in two or more languages. These words are potential cognates.
- Alignment: for each concept, the potential cognate words are aligned phone-by-phone, forming a list of potential sound correspondences.
- Evaluating correspondences: correspondences that occur in several words and help to exhaustively account for entire pairs of cognates are retained; words that cannot be explained by retained correspondences are rejected as non-cognate.
- Reconstruction of sounds: for each sound correspondence, the most likely proto-sound is hypothesized.
- Reconstruction of lexicon: reconstructed sounds are used to hypothesize how the sources of the cognate words were pronounced in the protolanguage.
- Subgrouping: correspondences are accounted for as the result of sound changes that split the protolanguage into a tree of intermediate protolanguages.
- Evaluation: tests are run to make sure all the posited rules work, ideally on new data as well as old.

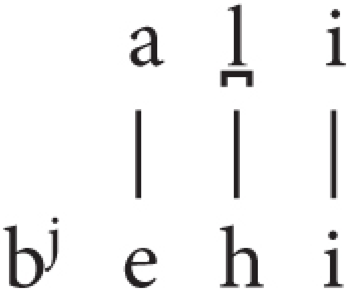
Stated as a sequence of simple steps, the comparative method looks like a natural candidate for computer implementation, and work toward that goal began as early as 1964 (Kay). As practicing linguists would attest, the problem is actually much more complex than this sketch suggests, and even today there exists no program into which linguists can pour their field notes and get back a full historical analysis of a language family. Instead, development has been proceeding along two tracks. One line of work set as its goal the development of aids for the historical linguist: the computer would do what it could do best, and its human assistant would do the rest. The other line of research concentrated on addressing the more complex problems, typically as an academic exercise in computational linguistics.

An early example of a tool that did a small part of the work well was the COMPASS program of Frantz (1970). Its main contribution was to tabulate the alignments hypothesized by a human linguist. The Electronic Neogrammarian of Hewson (1974) and the Reconstruction Engine of Lowe and Mazaudon (1994) did such things as generating reconstructions from attested words. Hewson (1993) found this helpful in finding cognates across Algonquian languages that had undergone semantic shift.

Another set of programs concentrated on helping with the evaluation phase of the work. Smith (1969) wrote a program to apply 21 rules, in sequence, to derive Russian words from Proto-Indo-European reconstructions. Similarly, Eastlack (1977) implemented (p. 135) the sound changes between Proto-Romance and Old Spanish as a suite of 42 ordered rules; given the Proto-Ibero-Romance reconstruction, the program would apply the rules and verify that the output matched the attested Old Spanish form. The Phono program of Hartman (1993) expressed the reconstructed segments by feature representations, allowing the researcher to write sound change rules at the featural level.

Most of the early programmers found that the stumbling block to fully implementing the comparative method lay very early in the procedure, in the alignment step. Kay (1964) wrote that the ideal procedure was to try all possible alignments and then solve for the ideal set of correspondences across the entire list of words. He then demonstrated that that procedure took several hours when used on just four pairs of CVC words. Kessler (1995) introduced the less ideal but much faster Levenshtein, or string-edit, criterion (Levenshtein 1965, Kruskal 1999). This criterion finds the cost of converting Word 1 into Word 2. The operations permitted are deleting a phoneme from Word 1, inserting into Word 2 a phoneme not found in Word 1, and substituting in Word 2 a phoneme different from that in Word 1. All of these operations have costs associated with them, typically fixed at 1 unit; matching identical phonemes in the two words has a cost of 0. The Levenshtein distance is defined to be the cheapest possible transformation. For example, when comparing [al̪ :i] and [bʲehi], two words for ‘cattle’, a lowest-cost transformation would be to insert a [bʲ] (cost 1), substitute [e] for [a] (cost 1), substitute [h] for [l̪ :] (cost 1), and match up [i] with [i] (cost 0), for a total cost of 3. Sounds associated by a substitution operation or by an identity match are considered a correspondence. When used on cognates, Levenshtein tends to give reasonable alignments, because on average, cognate sounds are more likely to be identical than are randomly matched sounds.

(1)



Many researchers have made modifications to the basic Levenshtein algorithm in order to more closely conform to intuitions about what sorts of sounds are most likely to correspond to each other. The earliest modification was to assign smaller costs to substitutions that are phonetically close to each other, which would seem advisable because cognate sounds tend to be phonetically closer to each other than to randomly selected sounds (Paul 1880). Adapting a technique introduced by Grimes & Agard (1959), Kessler (1995) described each phone as a bundle of 12 phonetic features, each of which took on numeric values. The difference between two phones, and thus the cost of aligning the one with the other, was the average of the corresponding values of each of the 12 features. Kondrak

(2000) reported good alignment of cognates in a model that incorporated properties such as the ability to ignore potential affixes; weighting features differentially such that some, perhaps place of articulation, are treated as more important than others; and incorporating one-to-many alignments, so as to better model breaking and fusion. Kondrak (2005) addressed the problem that Levenshtein does not take (p. 136) phonetic context into account by showing how it can treat sequences of segments as the basic units. Wieling, Prokić, & Nerbonne (2009) evaluated several variants, including one that permits crossing association lines, modelling metathesis. They reported especially good success with a version that adapts its substitution costs to favour the sound correspondences that are more frequent in the data.

The Levenshtein technique has been enormously popular in computational linguistics and many other fields that need to compare strings of symbols. Despite two decades of research, though, it is still not clear which variant works best for which problems. Indeed, researchers in several domains have reported that adding sophisticated phonetic information to the basic Levenshtein algorithm had little or no effect on accuracy in several domains (Kessler 1995 for dialectometry; Heeringa 2004, and Heeringa et al. 2006, for perceptual differences between words; Holman, Brown et al. 2011 for language cladistics). There are also other contenders for alignment procedures, including pair hidden Markov models (Mackay & Kondrak 2005), which take a long time to construct—they must be trained on large amounts of data—but perform very well once trained.

Subgrouping has been another aspect of the comparative method that has received a lot of attention from computational researchers, though not particularly often purely from the standpoint of phonology (see below under Cladistics for a brief discussion).

9.3 Language Relatedness

One important problem in historical linguistics is determining whether languages are related to each other. The comparative method is an excellent way of adducing evidence for relatedness. Unfortunately, almost any investigation that uses the comparative method turns up some evidence of language relatedness, even among unrelated languages, because of the factor of chance. Languages have large vocabularies but small phonological inventories, which means that some sound correspondences are likely to turn up if one looks hard enough. The upshot is that the results of most research, such as Vajda's (2010) fascinating proposal linking the Yeniseian languages of Asia and the Na-Dené languages of North America, are very difficult to evaluate, because opinions differ on how much evidence is enough (see, e.g. the critical review Campbell 2011). It would be useful if some statistical techniques could tell how to interpret the probativeness of a set of evidence. In most social sciences, researchers ask how likely it is that the data gathered in an experiment could be due to chance. If that probability is 0.05 or lower, most social scientists say it is statistically significant. Is such a thing possible in a historical social science, where experimentation is much more severely constrained?

9.3.1 The Significance of an Observation

The first mathematical technique developed to address this problem was probabilistic reasoning. Collinder (1947) decided that there must be a Uralic-Altaic family, because (p. 137) he found a set of 13 similarities shared by many of those languages. He reasoned that the odds of 13 similarities occurring by chance were vanishingly small. Hymes (1956) was more specific: Tlingit has a series of verb prefixes which come in the same order (aligned by function) as those typical of Athapaskan languages; the odds of that happening by chance are 12,168,189,440,000 to 1, therefore Tlingit is related to that family. Nichols (1996) calculated that any language for which the word for ‘widow’ has the consonants /w/, /j/, /d^h/, /w/, in that order, must be Indo-European, because the probability is 0.00000625, which, even when multiplied by the number of languages in the world, is less than 0.05.

Such demonstrations can be indicative, but not fully probative, for several reasons. First, linguists do not know enough to reliably calculate the probability that a particular event would occur in a language. This would either require uncannily precise knowledge of how languages are structured, or accurate counts based on surveying large numbers of languages that are known not to be related or to have influenced each other, not even indirectly. Second, even if the probabilities adduced were completely believable and accurate, it is very unclear how to interpret them. The reasoning behind such demonstrations is that the lower the probability of an observation, the less likely it could occur by chance. This is true enough, but in a very complicated system, as the world of language is, there will be many coincidences of very low probability. It is unclear how unexpected a single coincidence must be in order to definitively refute the idea that it is truly a coincidence; certainly that number is much, much smaller than 0.05.

In practice, science rarely proceeds by looking for amazing coincidences. If psychologists or shoe salesmen were looking to prove a relationship between IQ and shoe size, they wouldn’t scour the world for an individual with an extremely high IQ who wears extremely big shoes, and demonstrate how unlikely that individual is. The ideal situation is to have many observations, drawn in a way that is unbiased with respect to one’s research hypothesis. They might try to draw a random sample of a reasonably large number of people, at least several dozen, and measure their IQ and their shoes. With a large unbiased sample one can reasonably hope to compute statistical significance: how likely any correlation found is due to chance.

9.3.2 The Significance of Recurrent Sound Correspondences

Historical linguists search for recurrent sound correspondences between words of the same meaning because they believe that the association between sounds and meaning is arbitrary (Saussure 1916). If the same pair of segments repeatedly correspond when words of the same meaning are matched up, either a massive coincidence has happened or there is a historical connection between the languages.

Ross (1950) stated this conceptualization more precisely. Take two languages, say English and Latin. For a given concept, say 'thin', find the word that expresses it in English (*thin*) and note its initial phoneme, /θ/. Do the same for Latin *tenuis*, /t/. In a table whose rows are headed by English phonemes and whose columns are headed by (p. 138) Latin phonemes, add a tally mark for the /θ/≡/t/ cell (read: /θ/ corresponds to /t/). Pick another concept and repeat. After having sampled a large number of words, sum up the rows and columns. Now, in any cell, such as that for /θ/≡/t/, the number one expects to see, if the languages are not historically connected, would depend on those marginal counts. If out of 1000 words one had found 200 words beginning /θ/ in English and 150 words beginning /t/ in Latin, one would expect to find, by chance alone, (200 × 150) / 1000 = 30 tally marks in that /θ/≡/t/ cell. By the degree to which the actually observed tally exceeds the expectation, one would lean toward accepting that English and Latin may be historically connected, i.e. descended from the same parent language in a way such that the same proto-sound became /θ/ in English and /t/ in Latin. Ross reasoned that if one selects the concepts in an unbiased way—crucially one would not toss out the word *think* just because one knows in advance that *cogito* does not begin with a /t/—the cells of the table constitute a fairly large number of observations over which one can evaluate the null hypothesis. Over the table as a whole, are the tallies in the cells skewed so much from their expected counts that one can feel comfortable rejecting the default assumption of chance?

Unfortunately, reformulating the question in this way did not immediately lead to an answer. The ordinary test for association in a contingency table is a chi-squared test, which typically requires quite a bit more data than one is likely to have. Ross's procedure ends up with many cells whose expected counts would be very small, which would result in many false negatives if a standard chi-squared test is used. The ideal solution is a cumulative application of the hypergeometric distribution, but that takes so long to compute that even in the computer age researchers look for ways to avoid it (Wu 1993). Ringe (1992, 1993) used binomial approximations to the hypergeometric, a tack not uncommon among statisticians (Wu 1993), but Baxter & Manaster Ramer (1996) showed that there are realistic scenarios under which such approximations can lead to false positives.

Other workarounds look at ways of making the table less sparse so that easy statistics such as chi-squared can be used. Villemin (1983) reasoned that one can determine whether any particular sound correspondence such as /θ/≡/t/ recurs significantly more often than expected by reducing the entire table to a 2 × 2 table, so that the rows might be /θ/ and not-/θ/ and the columns might be /t/ and not-/t/. That would have worked if Villemin had stuck with that one test, but he independently tested each cell in the same way. As Ringe (1992) pointed out, if a table has 400 cells, that would mean running 400 separate tests. But if one asks 400 times whether correspondence counts have less than a 5 percent probability of occurring by chance, it stands to reason that about 5 percent of those tests will yield a positive result by chance. That is, such an approach is almost guaranteed to conclude that any two languages are historically connected.

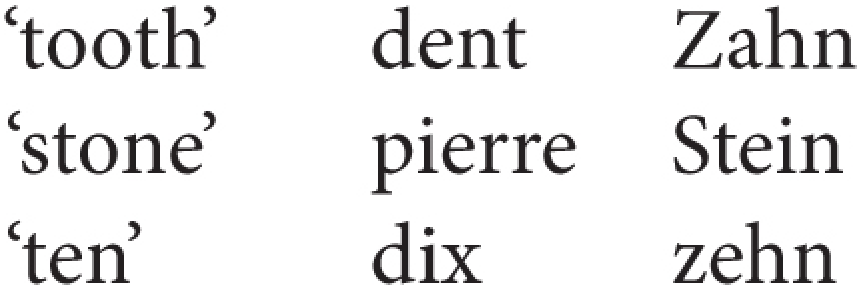
A statistically more valid approach to making the values in tables less sparse is to ignore certain phonetic contrasts. One tack is to have a single row and column for all vowels (Ringe 1992). This makes the table smaller and therefore the tallies in the cells bigger, making a chi-squared test more valid. Ignoring phonetic differences within a (p. 139) certain class of segments could be especially beneficial if one believes the distinctions within the segment type are particularly prone to loss or irregular changes, as vowels are in many languages. Other types of detail that might be worth ignoring in this way include phonation or manner of articulation. Kessler (2001) reported reasonable results when lumping together all consonants with the same place of articulation.

Ross's original (1950) idea was to make tables denser by using a thousand pairs of words. It may seem surprising then that linguists rarely use more than one or two hundred words (Villemin 1983 used 215 words; Ringe 1992 preferred 100), but there are several reasons for limiting the size of the word list. The most crucial reason is that each pair of words must be independent with respect to the arbitrariness hypothesis on which the test is founded. Consider what would happen if one used an entire English–Finnish bilingual dictionary: one would find, among others, a massively recurring correspondence between word-initial /ʌ/ and /e/. That sounds very impressive, but once one realized that almost all of those correspondences occurred in words that started with *un-* in English and *epä-* in Finnish, it would be clear that one is mostly dealing with a single morpheme, whose derivatives should be counted collectively as a single instance of /ʌ/≡/e/. Thus it is clear that word lists must be trimmed so as to include no two words that start with the same morpheme. Furthermore, vocabulary that is exempt from Saussurian arbitrariness, such as onomatopoeia and sound-symbolic words, should also be excluded from the test. And it is also important to exclude loanwords from most studies. If one were trying to discover whether Turkish shared a common parent with Urdu, one would not want to include on the list the many loanwords from Persian into each language. Thus there are many exclusions that need to be made just to satisfy the logic of the test (Kessler 2001); feeding in entire dictionaries would not only invalidate the test, but would probably end up biasing most analyses toward concluding that languages are related.

A less obvious problem with including hundreds of words is that not all words are equally probative. After many centuries, related languages are more likely to still share reflexes of the same word for ‘eye’ than for ‘river’ (Oswalt 1975). A test should ideally use only concepts that share the same low lexical replacement rate across the languages of the world. Quite a few studies have investigated which concepts have the lowest replacement rates (Swadesh 1955, Dyen, James, & Cole 1967, Oswalt 1970, 1975, Kruskal, Dyen, & Black 1973, Lohr 1999, Holman, Wichmann, Brown, Velupillai, Müller, & Bakker 2008, Tadmor, Haspelmath, & Taylor 2010). Swadesh himself started with 225 concepts in 1950 and was down to 100 in 1955; more recently several linguists have used lists of only two or three dozen concepts (e.g. Dolgopolsky 1986, Starostin 1991, Baxter & Manaster Ramer 2000, Holman et al. 2008). Although what constitutes the perfect universal concept list, if there is such a thing, is not yet settled, it is clear that only a fraction of the concepts from the Swadesh lists have the stability to be useful in even moderately deep linguistic relationships. At best, adding less stable concepts to the test waters down the evidence for language relatedness because the words for those concepts are more likely to be unrelated; this makes the data in the sound-correspondence table look overall closer to chance than it otherwise would be. At worst, less stable (p. 140) concepts are more likely to be loanwords, which can do real damage if they slip past the tester. Indeed, McMahon & McMahon (2003) actually used some words from the Swadesh lists, which were once considered the gold standard for stable, basic vocabulary, to look for patterns contraindicative of common descent. In light of all these issues, it now appears inadvisable to fill out sound-correspondence tables by adding more concepts. If anything, the concept list should probably be radically reduced from the hundred or more concepts that have been used in the past (Damerau 1975, Villemin 1983, Ringe 1992, Kessler 2001).

One promising technique for assessing the statistical significance of sound-correspondence tables uses rearrangement techniques to estimate what chance levels of correspondence would be (Kessler 2001). By definition, a significance level *p* in a statistical test is the probability that the observed evidence, or even more evidence, would occur by chance. There are many possible ways to construe *evidence* in sound-correspondence tables. Perhaps the approach that best captures the spirit of the comparative method is R^2, the sum of the square of the number of recurrences in each cell. There are also different ways of construing what *by chance* means: How can one decide what R^2 would be if the two languages definitely were not related? The most straightforward way to operationalize that idea is to note that if the languages were not related, then, by the arbitrariness principle, any word (sound sequence) in either language could have any meaning; or, to put it another way, any word in the one language could in principle have the same meaning—pair up with—any word in the second language. For example, when comparing French with German, the attested data would show:

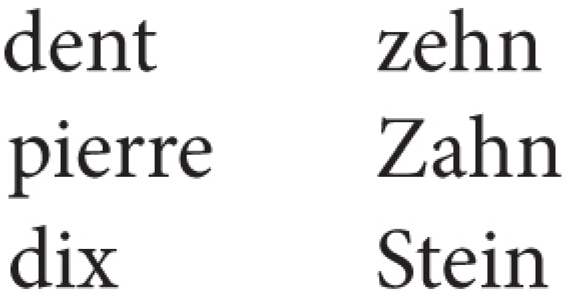
(2)



‘tooth’	dent	Zahn
‘stone’	pierre	Stein
‘ten’	dix	zehn

and one of six ways of rearranging that data fragment without regard to meaning would be

(3)



dent	zehn
pierre	Zahn
dix	Stein

Following through with this toy example, the observed data would yield an R^2 of 1 (2 instances, therefore 1 recurrence, of /d/≡/t͡s/, squared), whereas this chance configuration would yield an R^2 of 0 (no recurrences). Of course, this is only one possible way to rearrange the data. Ideally one would try all possible arrangements and see what proportion of them has an R^2 at least as high as the observed R^2. In this case one would get 2/6 = 0.33, which would be the *p* value. With larger word lists, where it would take too long to do all possible rearrangements, one can get a very good approximation of *p* by (p. 141) Monte Carlo techniques: computing R^2 over at least a thousand random rearrangements (Good 1994).

This first application, seeing whether the number of recurrent sound correspondences suffices to prove that two languages are related, has been described in some detail not just because it is a particularly long-lived topic of research, but also because it brings to the fore many techniques and questions that are good to keep in mind when considering many other types of computational historical phonology problems.

- The use of word lists is common in quantificational techniques. Many historical linguists have a reflexive aversion to them because they have been used as such mostly in glottochronology, which is widely discredited because its practitioners more often than not made naïve assumptions about lexical replacement rates. But it would be a mistake to equate all uses of word lists. In this specific case, it would be especially strange to be suspicious of the use of word lists inasmuch as the traditional comparative method has as its core the compilation of recurrent sound correspondences across words (Rankin 2005).
- The Swadesh concept lists are not sacrosanct. It often makes sense to use smaller lists selected primarily on the basis of their expected retention rates. However, this need must be balanced against the fact that a reasonable number of words is required to afford some statistical stability in the results.
- Any usefully sophisticated computational technique does not yield ‘the answer’. Instead of saying that Japanese is related to Korean, the computer estimates the probability that the evidence is due to chance.
- Reliable estimates of statistical significance entail feeding in data that is not preselected by whether it supports the researcher’s hypothesis.
- Lots of data is good, but uninformative data is bad. Computational methods need to find ways to privilege the most informative data. This was shown here of concepts that tend to have low retention rates across languages. Another example is that the initial phoneme of a word tends to be the most stable through time; throwing additional phonemes into the test just tends to water things down and make it harder to show that languages are related (Kessler 2001). This neglect of available data may seem undesirable, but statistical techniques are, after all, intended as an additional tool to supplement, not overthrow, other types of analyses.

Computers can automate many things, and indeed the rearrangement test described here relies crucially on delegating a lot of work to the computer. But linguists still have to work as hard as before to prepare their data for processing, such as by discarding words that violate the arbitrariness hypothesis. A recurring problem in quantitative linguistics is that people do not do this sort of preprocessing. They rarely say why, but a common belief seems to be that inaccuracy doesn’t matter: the law of big numbers means that all errors will cancel each other out. In reality, of course, only randomly distributed errors cancel each other out. Researchers who fail to preprocess their data often conclude that (p. 142) languages are related. Villemin (1983), for example, concluded that Japanese and Korean were related to each other, but regrettably had neglected to exclude from the data set words that both languages had borrowed from Chinese.

9.3.3 The Significance of Sound Similarity

Tabulating recurrent sound correspondences across word lists is essentially a phonological methodology, in that the point is to find how many sounds survive from a putative parent language. Nevertheless the technique described in the previous subsection does not really ask linguists to apply a lot of phonological knowledge, unless they choose the option of collapsing sets of phonologically similar sounds into one row or column. But Oswalt (1970) introduced a new way of testing whether languages are related: languages whose words for the same concept sound more similar than one would expect are diagnosed as being genetically related.

The very idea of using phonetic similarity sounds heretical, because historical linguists have always had to spend no small amount of time patiently explaining that Basque and Mayan are not necessarily related even though an enthusiast has found dozens of words with somewhat similar meaning that have somewhat similar sounds. The community is only just beginning to settle down from the furore (e.g. Campbell 1988, Matisoff 1990, Salmons 1992c, Ringe 1996a) over claims made by a highly respected linguist that eyeballing words for similarities across languages is a satisfactory way to prove that large numbers of languages are related (Greenberg 1971, 1987, 2002). But, notwithstanding the fact that many people have applied phonetic similarity in ways that can most charitably be described as irreproducible, the core of the theory is sound. As languages diverge, an original sound will either remain the same in all descendants or it will begin to split into two sounds, which usually remain rather similar in the first instance (Paul 1880). Thus the words for concepts in two daughter languages will be more similar than those in two unrelated languages, with the measurable similarity decreasing gradually over time.

The main issues are whether one can objectively measure phonetic similarity and figure out whether the amount measured is greater than chance levels. This latter problem can easily be handled in the same way it was handled above for sound-recurrency tables: rearrangement significance tests. The general notion of rearrangement statistics for phonetic similarity measures was introduced by Oswalt (1970), who compared the observed similarity statistic with the statistic he got when words were shifted 100 times by a constant factor. Baxter & Manaster Ramer (2000) recast this test as a Monte Carlo rearrangement test, using enough rearrangements to make the test truly informative.

Rearrangement tests give linguists a great deal of flexibility in designing tests for measuring the phonetic similarity or difference between words in two different languages. A useful convention, though, is to conceptualize the tests as metrics, or distance functions. That is, if two words are identical in all important ways, the measurement (p. 143) would be 0; words that are not identical would get a positive measurement; the more different they are, the more that measurement would differ from 0; and the number should be the same whether A is compared to B or vice versa. Oswalt (1970) used the simplest possible metric. For him, two words were either similar (0) or not (1). It is also possible to have multivalued or continuous metrics that express just how different the words are (Kessler & Lehtonen 2006, Kessler 2007).

An important question is what parts of the words should be compared. It is tempting to devise a metric that takes all the information in the two words into account—one hates to be wasteful with hard-won data—but significance tests have more power when they are fed only high-quality data. Some small-scale experiments (Kessler 2007) suggest that it is best to consider only the first phoneme of each root, comparing the two only with respect to place of articulation. This finding matches the rule of thumb that as words change over time, the place of articulation of the initial phoneme tends to be the most stable feature, and indeed most researchers simply worked with this feature without further ado. Oswalt (1970), for example, experimented with different metrics, but their one constant was that words that begin with different places of articulation had to be considered dissimilar (distance 1). Baxter & Manaster Ramer (2000) got at the same thing by classifying words as similar if and only if their initial phonemes fell in the same Dolgopolsky class. Dolgopolsky (1986) had set up 10 mutually exclusive equivalence classes for sounds, grouping together those sounds that are more likely to correspond in cognates in the Eurasian languages he investigated. For the most part these classes were characterized by place of articulation. Turchin, Peiros, and Gell-Mann (2010) also used Dolgopolsky classes and rearrangement significance tests (the bootstrap method) in their testing of the Altaic hypothesis.

Currently, there is little evidence that one metric is significantly better than another. The obvious way to calibrate and evaluate such metrics would be to take a large random sample of languages and show that one metric is better than another at giving positive results for languages known to be related and negative results for languages known to be unrelated. An inconvenient fact standing in the way of doing that is that no pair of natural languages is known a priori to be unrelated—indeed the monogenesis of all languages is a viable theory. If one metric reveals a weak connection between Indo-European and Uralic, and another does not, there is no gold standard for deciding which is correct (Kessler 2007). Computational techniques are subject to the same epistemological asymmetry as the traditional comparative method: it is possible to give a convincing demonstration that two languages are related to each other, but it is not possible to demonstrate that they are not related to each other.

Databases are now available with substantial lexical information on a great many of the languages of the world. It is tempting to process them all in order to find as many relatives as possible. But caution is in order. The data must be carefully processed to omit onomatopoeia, loanwords, and redundant morphemes, a task that may prove difficult for thousands of ill-attested languages. Even more importantly, one expects to get lots (p. 144) of false positives when lots of tests are run: approximately at the same proportion as the chosen significance cut off, which is typically 5 percent. Massive searches for significant data require techniques such as false discovery rate analysis (Benjamini & Hochberg 1995) to help limit the reporting of false positives.

9.4 Phonetic Phenetics

In phenetics, the goal is to classify objects on the basis of their synchronic similarity. A phenetic analysis is often the first step in a historical analysis. Because phenetics tends to be simpler than a full-scale historical analysis, it can provide a quick lay of the land, helping the linguist generate hypotheses. Similar languages are more likely to be in the same family or branch than dissimilar languages; similar dialects are more likely to have shared more history than dissimilar ones. There can also be an advantage in the relative theory-neutrality of phenetics. Purely historical analyses, especially if they are driven by a very specific model of language change, may yield incomplete or confusing results that a linguist might be able to better interpret with the aid of good analyses and visualizations of the phenetic landscape.

One of the first applications of a truly quantitative phenetics in linguistics was dialectometry (Séguy 1971), which produced rich visualizations of the similarities between dialects of the same language. The earliest studies emphasized lexical variation, but later studies incorporated some phonological issues; for example, Babitch and Lebrun (1989) classified Acadian French dialects by their realization of /r/. Kessler (1995) introduced a purely phonetic phenetic procedure, comparing the full pronunciations of words that share the same gloss in different Gaelic dialects (Wagner 1958). The phonetic distance between each pair of dialects was taken by summing the distance between the words for each of 51 concepts. The distance between two matching words was defined in terms of the Levenshtein distance between them.

Of course, there are many different linguistically informed ways to compare phones, and several researchers have come up with different schemes (for a fuller review, see Kessler 2005). One improvement has been to give greater weight to some features than others (Juola 1996, Kondrak 2002). Another approach has been to back off from features entirely, on the theory that the important differences between sounds is not simply a linear function of their individual features taken independently (but cf. Nerbonne & Heeringa 1997, who found features advantageous). Oakes (2000) gave substitutions one of two different weights, based on whether they are an example of a well-known type of sound change. Heeringa (2004) tried using acoustic features, including a comparison between the spectrograms of the sounds in question.

In order to better conceptualize the patterns between what may be thousands of pairs of dialects or languages, it is convenient to group them by how similar they are. Babitch and Lebrun (1989) used UPGMA (unweighted pair group method with arithmetic mean: Michener & Sokal 1957), a simple agglomerative clustering technique that (p. 145) is often effective in phenetic studies. They went through the entire distance matrix and found the dialects that have the smallest phenetic distance between them. Those dialects were proclaimed to be a dialect group, which then replaced its two members in the distance matrix; the distance between that group and all the other dialects were treated as the average of the distance from each of its members. This process then continued iteratively until there was only one big dialect left. The chain of dialects within dialects can be expressed as a binary tree or as a box diagram in the manner of Dyen, Kruskal, and Black (1992). Kessler (1995) tried UPGMA clustering for Gaelic in order to verify that the groups that emerged were the same as those found by dialectologists, and they were.

There are many alternatives to UPGMA clustering as a final step, such as neighbour joining (Saitou & Nei 1987), cluster analysis, and multidimensional scaling (Kruskal & Wish 1978, Heeringa 2004). As for many other techniques in computational linguistics, these new methods in phenetic analysis are often borrowed from other disciplines. Felsenstein (2008) described three programs for evaluating distance matrices in biology—Fitch, Kitsch, and Neighbor—which were developed as part of the PHYLIP package for phylogenetic research in biology (Felsenstein 2009). SplitsTree (Huson 1998, Huson & Bryant 2006) incorporates many different algorithms for processing distance matrices, including those allowing them to be visualized as networks. Networks are much like trees, except that inconsistencies in the data are explicitly represented by branches that connect ordinary splitting branches. McMahon & McMahon (2005) present a good introduction to the use of these techniques for historical linguistics.

McMahon, Heggarty, McMahon, & Maguire (2007) used SplitsTree to study phonetic similarity between English dialects. Instead of a concept list, they began with a list of 60 reconstructed Proto-Germanic words and compared the reflexes in 19 descendant language varieties, mostly accents of English. Thus German *Hund* was compared with its English cognate *hound*, not with the synonymous *dog*. Instead of using the automatic Levenshtein procedure, the sounds of each word were hand-aligned to their Proto-Germanic ancestor, so that the programs could compute the distances between sounds known to be cognate. Distances between pairs of sounds were performed primarily on the basis of phonetic features. The distance information was fed into SplitsTree, using two algorithms: Neighbor-Joining, which draws trees, and NeighborNet, which draws networks. For both algorithms, fairly reasonable representations emerged, with much reticulation in the latter case. Interesting results were explored by plotting the distance between three localities on each of the 60 cognate sets. For instance, General American English was shown to be closer to Scottish and Irish English than to the dialects of England it derives from historically. By plotting the distances each Glasgow pronunciation is from both General American and English Received Pronunciation, the researchers found the words where the Glasgow pronunciation was closer to the former than to the latter: most were words like *year, four, fire*, and *horn*, revealing that General American clusters with Scottish because of their shared retention of the rhotic gesture in syllable codas.

(p. 146) A good deal of research is continuing into phonetic phenetics. One of the more difficult outstanding problems is that it is unclear how to measure precisely how successful any specific methodology is: is there a gold standard that says how similar two languages or lects are? Heeringa and colleagues (2006) may have come the closest to this ideal, when they correlated their various distance scoring methods with scores obtained by asking humans to rate how similar utterances were. But phenetic analysis may be used not just when researchers are directly interested in similarity between language varieties, but also when they are preparing the groundwork for a historical analysis. In such cases, it may make sense to fine-tune the methodology so that the distances it yields correlate with known linguistic connections rather than perceptual judgements.

9.5 Cladistics

Cladistics is the evolutionary development of a language family. Cladistics can be thought of as classical subgrouping, at least as the latter is most rigorously defined. The definition excludes groups that are defined solely along phenetic, geographic, or cultural lines.

Despite the strict distinction between cladistics and phenetics, it is common in both biology and linguistics to use a phenetic analysis as an approximation to a cladistic analysis. Phenetic analyses tend to take only seconds of computer time; full cladistic analyses may take days or weeks. A significant problem, though, is that there are plenty of well known, commonly occurring diachronic situations that lead to conflicts between cladistics and phenetics. The most serious issue is illustrated by the aforementioned McMahon et al. (2007) results that suggested grouping General American with the other rhotic accents. It may well be true that to some ears, at least, General American sounds more Scottish than English on account of its many /r/ sounds. However, in this case the rhotic accents share a retention (symplesiomorphy) from early Modern English, not an innovation (synapomorphy), and so grouping them together on that basis would be cladistically incorrect. Another problem is that individual historical events can yield massively different distance measures. A single apocope event in one language could make all of its words substantially different from the words in another branch; whereas a series of conditioned changes affecting only a few features in a few words may result in negligible effects on overall vocabulary distance. Or a change affecting a frequent phone like /r/ would have a more profound effect on the distance matrices than multiple changes affecting an infrequent phone like /v/. To this must be added the usual diachronic bugaboos: consider the effect on distance matrices of a sound change followed by its reversal ([t] > [θ] > [t]), identical sound changes occurring independently, and, of course, borrowings. Users of phenetic models must constantly keep in mind the fact that they are doing something that is fundamentally different from classical phylogenetic reasoning.

(p. 147) Is it possible that computer programs can be turned loose on large databases and tasked with producing a cladogram for all the languages of the world? The Automated Similarity Judgment Program project is doing something very much like that, classifying thousands of languages and dialects by their lexical similarity to each other (Brown, Holman, Wichmann, & Velupillai 2008, Müller et al. 2010). As the ASJP researchers make clear, however, several cautions are in order if one is looking specifically for a phylogenetic analysis. All of the warnings given earlier still apply, including the need to preprocess data and the general caution that phenetic analyses may correlate strongly with phylogenetics, but they may also diverge spectacularly in some cases. In addition, it needs to be noted that finding a measurable amount of similarity between languages is not the same as proving they are related, until one demonstrates that the similarity exceeds chance levels. The fact that phenetic trees give reasonable results for languages known to be related invites the inference that the nodes they draw at higher levels are equally reliable, when it fact they may be based largely on similarities due to chance.

There also exist computer programs for performing cladistic analyses in ways close to traditional manual subgrouping. Programs that have been used by linguists include MrBayes (Ronquist, Huelsenbeck, & van der Mark 2005) and PAUP* (Swofford 2007). The linguist identifies phylogenetic characters; for historical phonologists, these might be sound correspondences. A character matrix is set up, showing which reflex each language has. The computer program then attempts to find the tree or trees that can account for that character matrix while optimizing certain conditions. For example, the program may try to set up branching so as to minimize the number of sound changes.

I will not go into much detail about cladistic programs here, mostly because their use in phonology is still highly experimental. Most of the best known cladistic analyses in linguistics have used few if any phonological characters. Of the 376 phylogenetic characters used by Ringe, Warnow, and Taylor in their Indo-European classification (2002), only 22 were phonological. More commonly, linguists use only lexical data (e.g. Gray & Atkinson 2003) or typological data (e.g. Dunn, Levinson, Lindstrom, Reesink, & Terrill 2008), or both (e.g. Wichmann & Saunders 2007). There is no real state of the art for phonology-based cladistics, so the reader is directed to more general treatments of linguistic cladistics. McMahon & McMahon (2005) have given a gentle introduction to the field, and Nichols & Warnow (2008) have supplied a thorough tutorial and review of the literature, including some methods I am here characterizing as phenetic.

Using computers for cladistics is clearly the wave of the future, but the present is still unsettled. Experiments such as those of Nakhleh, Warnow, Ringe, & Evans (2005) and reviews such as Nichols & Warnow (2008) have demonstrated that varying the methodologies very often leads to incompatible results. Another difficulty is that coming up with the best tree to fit the data is ridiculously time-consuming, because the number of possible trees that can be drawn increases superexponentially with the number of languages. Even the heuristics designed to make a best guess without actually inspecting (p. 148) all the alternatives may take days or weeks to run, and it is never quite clear when they are finished. In addition to these problems, there remain, as always, all the familiar problems from traditional subgrouping, such as deciding on the probability that sound change may repeat independently. Nevertheless, computational cladistics holds great promise for linguistics, and has been perhaps the most fervent area of computational linguistic research of recent years.

Brett Kessler

Brett Kessler is an associate professor at Washington University in St. Louis, where he teaches in the Linguistics Program and the Philosophy-Neuroscience-Psychology Program. He works on developing computational techniques for studying language phylogenetics and the psychology of phonemic writing systems, with emphasis on statistical methods for hypothesis testing in linguistics.