Transactions of the Philological Society Volume 103:2 (2005) 243–260

PHONETIC COMPARISON ALGORITHMS[1]

By Brett Kessler
Washington University in St. Louis

Abstract

Appealing to phonetic similarity has traditionally been discouraged in linguistics, partly because it has been an ill-defined and subjective concept. But much research nowadays requires measures of similarity between words, from practical work in speech technology, information retrieval, and commercial branding to theoretical studies involving language comparison and history. Phonetic comparison algorithms are crucial to this work, enabling computer implementation as well as reliability and significance testing. But phonetic similarity is not a unitary concept. Various types of measures are discussed, with emphasis on those most appropriate for current and future work in historical linguistics.

1. Introduction

Phonetic comparison algorithms are precisely defined methods for quantifying the similarity between speech forms — segments, words, or even entire languages — on the basis of their sounds. People outside the field of linguistics express polite amazement when told that linguistics offers no standard way to define how similar two words are. There are good reasons for this deficit: mostly, linguistics lacks such a measurement for want of caring. Mainstream linguistics has always embraced mathematics that are

[1]This review conflates phonetic and phonemic comparison in that most of the methodologies surveyed treat phonemes as idealized phonetic segments between which some phonetic or featural distance is to be computed. Regrettably I must leave unmentioned a large literature that compares phonemes or entire phonological systems on other bases (e.g. Altmann & Lehfeldt 1980).

categorical rather than continuous (Joos 1950). In phonology, a segment either meets the condition for a rule or constraint, or it does not. In historical linguistics, recurrence of sound correspondences is much more important for establishing cognacy than is degree of phonetic similarity.

This emphasis on categories over quantities is certainly justified. A sound that meets two thirds of the conditions for a sound change does not undergo two thirds of a change. Synchronically, language competence is basically categorical. Humans find it easy to make binary judgements about linguistic entities, but are typically inconsistent in making more quantificational judgements (Zobel & Dart 1996) or even relative rankings ("Is [dɔg] more like [dɔn] or like [lɔg]?") when grammaticality or communicative content is not at stake. Because so much about linguistic competence is clearly categorical, and because people's quantificational judgements about similarity are so variable, it should come as no surprise that formal linguistics offers no precise rules for measuring similarity.

But similarity was never completely banished from linguistics. When the idea of phonemes was popular, it was always held that allophones must have some similarity to each other, lest we have to say that pairs like [h] and [ŋ] are allophones in complementary distribution. Similarly, sound change was generally considered to be gradual (e.g. Paul 1880), which can only mean that the output is similar to the input. Perhaps most importantly, it has always been noted that sound change and phonological constraints often operate not on a single specific segment but on a set of similar segments, such as all voiceless stops. Even if similarity does not apply to language competence, it may be very useful for understanding perturbations in performance, and therefore linguistic change and its outcomes.

Indeed it turns out that there are many questions for which people have required some sort of standard for measuring the similarity or difference between words. Here is a very incomplete list for which space is lacking for all but the most rudimentary references. I start with some applications that may seem remote from theoretical and historical linguistics but which contain ideas that may be applicable more broadly.

2. Applications

Speech applications in effect measure the phonetic distance between a given utterance and reference words. Such measures are used to identify spoken words, to assess second language proficiency (Bernstein, Barbier, Rosenfeld & De Jong 2004), to diagnose articulatory problems (Connolly 1997), to quantify children's acquisition of pronunciation (Somers 1999), to identify languages (Muthusamy & Spitz 1997), and to verify the identity of a speaker (Furui 1997). Speech technology may seem a world apart from linguistics proper, but many linguists have experimented with using acoustic measurements as a more faithful and direct representation of speech.

In commercial branding, developers of trademarks seek to avoid creating names that sound confusingly similar to any of the thousands of other trademarks. In areas such as pharmaceuticals, phonetic similarity can even have serious health effects (Kondrak & Dorr 2004; Lambert, Chang & Lin 2001). Auditory distinctiveness is also paramount in developing oral security and encryption keys (Juola 1996).

One approach to spelling correction is to determine what sequence of sounds a given spelling could represent then return the set of real words that are phonetically closest to that sequence (Toutanova & Moore 2002). Analogous techniques can be used in cross-language information retrieval (Fujii & Ishikawa 2004). People who enter the term *collocation* into a search engine could retrieve Japanese papers that contain the phonetically similar loanword *korokeisyon*.

In computational linguistics, the automatic alignment of multilingual text can be the first step toward such goals as machine translation and the production of bilingual lexicons (Hiemstra 1998). Words that are especially similar to each other can be used as islands of certainty in the automatic alignment process (Kondrak, Marcu & Knight 2003).

There are also several applications in the field of historical linguistics. Kondrak (2002) developed techniques for identifying likely cognates through phonetic similarity. It is natural to object that there are cases where cognates are not similar to each other and

cases where similar words are not cognate. But like many new computational techniques, the task is conceived as probabilistic: one looks for answers that are correct within a specific confidence interval. Such fuzziness is often more honest than proclaiming complete certainty and may be less damaging to the discipline than demurring from posing questions that cannot be answered with certainty (Embleton 2000). In any event, if one has to find possible cognates in the absence of semantic glosses, matching words by phonetic similarity may be the only reasonable first step (Melamed 1999).

A further use of phonetic similarity measures is to align segments that correspond to each other in two variants of the same word. Somers (1998) aligned young children's speech productions with the presumed target; for example, if a child says /fi/ for/θri/,/ f/ would be aligned with intended /θ/, /i/ with /i/, and /r/ would be unaligned. Oakes (2000) automatically aligned cognate pairs by phonetic similarity and considered segment pairs that aligned more than once to be recurrent correspondences.

Some lines of research use phonetic similarity to prove that languages are historically connected. Much research along these lines has been roundly criticized (e.g. Matisoff 1990), but arguably the main problem has been the lack of statistical controls, including a precisely defined phonetic comparison algorithm (Kessler & Lehtonen, in press). Oswalt (1970) was apparently the first to introduce well defined phonetic similarity metrics, although the statistical test to utilize them was not perfected until later (Baxter & Manaster Ramer 2000).

In other research, the immediate goal in comparing languages is to draw a map of the linguistic landscape for a continuum of languages or dialects. Such maps may enable one to draw implications about past and present social connections between regions. Understanding the objective relations between dialects also lets one test hypotheses about the cultural construction of perceived relations between speech varieties or the relation of objective differences to interintelligibility. Dialectometry arose out of such a desire to quantify the relations between dialects (Séguy 1971), but the bulk of interest was initially in lexical variation. Phonetic comparisons were typically restricted to a few specific phonemes

that had previously attracted attention (e.g. Babitch & Lebrun 1989). Kessler (1995) introduced the idea of quantifying the phonetic distance between pairs of words. His intention was to express, to the intent possible with one crude number, dialect differences resulting not only from sound change but also from lexical replacement. Juola (1998) went one step further by comparing languages based on whole texts. He used traditional orthography, but his technique could easily be adapted to phonetic transcriptions.

Lastly, some work has used phonetic comparison to determine the time of separation of languages or their subgrouping (cladistics). The general approach is contentious, in part because of its association with impressionistic methodologies such as multilateral comparison, but the situation is improved when a precise algorithm for phonetic comparison is used. Grimes (Grimes & Agard 1959; Grimes 1964) did so in determining how far sounds have changed in cognate words in Romance. Heggarty (2000) aggregated such measurements over 40 cognates to compute similarity for all pairwise combinations of several Romance, Germanic, and Slavic languages. The expected groupings emerged, with only French proving a bit recalcitrant. But Heggarty also showed that the Romance languages vary widely in their similarity to their parent Latin. He concluded that there is such great variation in rate of phonetic change that similarity-based (phenetic) language groupings should be used to infer cladistics only with proper appreciation of the very large confidence intervals involved.

3. Techniques

As the diversity of the above list of applications suggests, phonetic distance can mean quite a few different things: Difference between acoustic properties of the speech stream; difference between articulatory gestures; perceptual distance between isolated sounds; judged distance between sounds in communicative context; or historical distance between sounds, in time or in number of events. Choosing which type to use in a particular study is not always straightforward, nor is the choice often explicitly justified. Articulatory gestures seem to be used especially often because historical

distances are unavailable, or because they are easier to obtain or seem more objective than human judgements.

Most techniques for measuring phonetic distance take the distance between phonological segments as point of departure. The simplest of these approaches is to recognize two distances: one for identical sounds (typically 0) and another for nonidentical sounds (1). Such a simple, binary, technique often performs surprisingly well in assessments. Kessler (1995) found that his dialectometrical analyses based on binary measures were more congruent with traditional isoglosses than analyses using more complicated measures. Heeringa (2004) found that binary distances resulted in dialect difference measurements that were the most congruent with the judgements of native speakers of those dialects. Perhaps the embarrassingly simple binary methods worked well because the applications were fundamentally binary. It is obviously the case that isoglosses are binary. As for Heeringa's human judgements, they were multivalued, but human judgements may be most definite when dissimilarity passes a functional or sociolinguistic threshold, resulting in confusions between words or in dialectal shibboleths. If an application models inherently binary processes such as those that may be involved in certain types of human perceptions, utterly basic binary techniques may actually be appropriate.

Binary comparison tends to hide the fact that somebody has to decide what variations between phone segments are to be ignored when deciding whether two sounds are instances of one idealized phone type. This task can be quite vexing cross-linguistically, when one cannot resort to principles of phonemic theory such as complementary distribution within the vocabulary of a single language. A more generalized and sophisticated approach is to explicitly divide the phonetic inventory into equivalence classes. Dolgopolsky (1986) defined 10 sets of consonants, such that consonants were deemed more likely to change into consonants within their own group than into consonants in another group. For example, one group consisted of labial obstruents, another of nonsibilant coronal obstruents. Effectively, a pair of sounds gets difference score 0 if both are in the same group and score 1 if they are not. Oswalt (1970, 1991) used even more complicated definitions

of binary equivalence classes, one of which states that consonants are in the same equivalence group if they have the same point of articulation and agree in voicing, stoppage, and nasality, or two out of the three. The metric is still binary, but the lines are drawn at a different place.

Perhaps the most common way to measure phonetic similarity is to compute over feature bundles. Grimes & Agard (1959; Grimes 1964) compared the linguistic difference between Romance languages by looking at sound correspondence sets, one set for each sound change. Each phonetic segment was represented as a set of six multivalued articulatory features following Pike (1943). Each of the features took numeric values, whose rank but not magnitude followed a natural order. For example, the point of articulation feature had seven possible values: 1 for bilabial, 2 for labiodental, up to 7 for glottal. The distance between two sounds on any feature was the magnitude of the difference between those numeric codes, and the distance between two sounds as a whole was the sum of those differences. The measurement was based on the idea that sound change is incremental, so that if, say, a change from a velar like /k/ (6) to a palatal like /c/ (5) is one unit of change, a change from /k/ to bilabial /p/ (1) represents six times as much change.

The phonetic distance metric used by Grimes & Agard (1959) was not much different in spirit from distance metrics used up to the present day. But they all differ somewhat in detail, and the differences can substantially affect the outcome of a study.

Feature sets have been based on many different systems. The most interesting contrast is between those using perceptual features and those using articulatory features. The few direct comparisons that have been made give the nod to the latter, although the question is far from settled. Somers (1998) compared two types of feature systems discussed by Connolly (1997), to see which would better align children's imperfect speech productions with the presumed target. Somers gave no numbers, but there is no reason to doubt his conclusion that the articulatory approach gave more intuitively correct alignments than a simple acoustic approach. Heeringa (2004) found a similar result even with a task that was clearly perceptual, the judging of dialect differences.

The feature values used are typically either binary or multiple-valued. The binary approach has the advantage of congruence with theoretical linguistics, which emphasizes categorical distinctions. The multivalued approach agrees more with the continuous mathematics used in phonetics and maps more naturally to the concept of distance. Kondrak (2002), for example, used a system based on Ladefoged (1995), where a feature called Place can take on 11 values ranging from bilabial (1.0) to glottal (0.1), and all values in between roughly reflect actual distances in the mouth. In one of the few published comparisons of a binary feature system (Hoppenbrouwers & Hoppenbrouwers 2001) with a multivalued system (one based on Almeida & Braun 1986), the former correlated with perceptions of dialect differences better than did the latter (Heeringa 2004). But the Almeida & Braun scheme, like that of Grimes & Agard (1959), assigned feature values on an ordinal rather than a continuous scale. It isn't a very promising starting point for an algorithm for computing distances if one cannot subtract two values for a given feature and obtain a meaningful distance measure.

Comparing the distance between two segments across multiple features is even harder, in that standard phonological theory offers no guidance how to do so. A popular approach is the Manhattan distance: compute the distance between the segments based on one feature at a time then add up those distances across all the features. Grimes & Agard (1959) justified such an approach by appealing to Austin's conjecture (1957) that sounds change in place or in manner, but not in both directions simultaneously; therefore if segments differ in both of those two features, it must be the result of two independent, therefore additive, events. Other summarizing functions have been tried, including the Euclidean distance, where one squares the differences between the feature values before summing them, then takes the square root of that grand sum. Heeringa (2004) found this to be slightly more effective than the Manhattan distance. If this result holds, it suggests that, at least in the domain of dialectometry, Austin's conjecture is not completely applicable to all features.

Occasionally researchers weight the different features to reflect their differential saliency. Such an approach is implicit in the

scheme of Juola (1996), who represented sound segments as a vector of bits and computed the distance between two segments by counting the number of bits that differed between them. Crucially, he assigned, for example, more bits to the representation of voicing, which he considered to be highly salient, than to features like nasality, which he considered less salient. Kondrak (2002) made the weightings more explicit and flexible, by expressing them as coefficients that could be easily changed to any numeric value. He was able to exploit this flexibility by adjusting the coefficients until he got optimal performance on aligning cognate words.

Though this use of weightings is advantageous, it still leaves us with a linear, additive model of feature effects, where it is assumed that the contribution of each feature is independent of the contribution of any other feature. But it may be that place of articulation, for example, will be very salient for English obstruents, but not so much for nasals. Not much attention has been paid to this issue as a mathematical problem, but some researchers have tried approaches that address this interaction indirectly. One approach that has a certain elegance is to reduce the problem by ignoring most features. Covington (1997) had reasonably good success aligning cognates by considering little more than whether the segments were consonantal or syllabic; interactions between his small set of features were handled by spelling out the distance to be ascribed to each possible combination of feature values. Kessler & Lehtonen (in press), in computing the probability that languages are related, ignored all but the place of articulation. This was done because place is the most durable feature over time, but a welcome side effect was that all distances between phonological segments could be expressed simply as the distance between their places of articulation. A more sophisticated approach, by Oakes (2000), considered whether substituting one segment for the other constituted a well known type of sound change; such pairs (e.g. /f/:/h/, lenition) were assigned a distance of 1 from each other, while other nonmatching segments (e.g. /s/:/k/) were given a distance of 2. It is disappointing that the algorithm performed fairly poorly, especially on language pairs that are rather remotely related (Kondrak 2002). A factor contributing to its relatively poor performance — besides the fact that its competitors had been

tuned for performance on the evaluation suite — is that all sound changes were given the same score regardless of their likelihood.

Another solution to the problem of interaction between features is not to use features at all. Heeringa (2004) reported that in his perceptual task, measuring the difference between spectograms of reference sound segments worked about a percentage point better than feature-based comparison. Another approach that has rarely if ever been used in serious applications is to gather individual measurements for each pair of segments in a way relevant to the task. For example, if the application or theory being tested is essentially perceptual in nature, one could get people to judge the similarity or difference between each pair of sounds. No matter how elegant and convenient mathematical approximations may be, it is always wise to base the distance metrics on as much empirical data as possible. It may be tedious to collect hundreds or thousands of judgements, but often the results are reusable, and a good start has been made toward collecting and sharing such data (e.g. Miller & Nicely 1955; Singh & Woods 1971; Singh, Woods & Becker 1972).

In some applications, it is desirable to let a single segment stand in for the entire word. When investigating whether languages are historically connected, the main approach has always been to compare just the first segment, or perhaps the first consonant, of the words in question (e.g. Oswalt 1998), because in almost all cases, the rest of the word will be less probative and only dilute the evidence. But in many other applications, the whole point is to compare entire words. When creating brand names, one must ensure that the name as a whole does not sound too much like the name of a competing product (Lambert, Chang & Lin 2001; Kondrak & Dorr 2004); when measuring the differences between languages or dialects, one might want to include as much information as possible in one measure (Kessler 1995). A favoured method for comparing whole words is to compute the Levenshtein, or string-edit, distance between the two words (Levenshtein 1966). This measure involves the optimal alignment of words. Costs are assigned to each of the individual segments that don't match up; they are called deletions or insertions. Costs are also assigned to pairs that do match up; these are called substitutions. There may be many, perhaps thousands, of possible alignments; the Levenshtein

distance is the sum of the insertion, deletion, and substitution costs of the alignment that gives the lowest sum. Dynamic programming algorithms exist that are much faster than exhaustively looking at every possible alignment, so the Levenshtein distance can be efficiently computed for even quite long words (Kruskal 1999).

The basic version of the Levenshtein measure is binary: it assigns uniform costs (1) for all insertions and deletions and for all substitutions that do not involve a pair of identical segments; matches of identical segments always have a cost of 0. More commonly, people prefer to assign different costs to different insertions, deletions, and substitutions. These may reflect the acoustic salience of a certain type of sound or the phonetic distance between two sounds involved in a substitution, or perhaps, as in Oakes (2000), an estimate of the likelihood that some sound change would result in the substitution, insertion, or deletion in question. Variants of the Levenshtein measure have been developed to handle special cases useful in linguistics. Kondrak (2002), for example, favoured local alignment when matching cognates; this is a variant that can ignore material at the beginning and end of words and so is useful in situations where words may share the same root but have additional morphemes that differ. There are versions that can handle many-to-one alignments (e.g. breaking and fusion, also discussed by Kondrak) and transposition (metathesis). However, one major shortcoming that is rarely discussed is that the phonetic environment of the sounds in question cannot be taken into account, while still making use of the efficient dynamic programming algorithm. Oakes showed that processes like assimilation can sometimes be modelled by treating e.g. /n/ $\rightarrow$ /m/ before /p/ as if it were the simple replacement /np/ $\rightarrow$ /mp/, but it is easy to see how this sort of solution could become unwieldy if not unworkable. In cases of long distance assimilations, for example, such as where the first vowel of the word affects the backness of all other vowels in the word, one would need a separate replacement rule for virtually every possible word, which amounts to having no phonetic comparison algorithm at all. Currently, the predominant solution to this problem is to ignore context entirely. Another solution is not to rely on dynamic programming. Covington (1997) used exhaustive search in his cognate alignment program, which was perhaps

unnecessarily inefficient for his purposes, but such an approach would allow environment to be taken into account. While speed is very important in some applications, such as real-time speech recognition, historical linguists may well have the patience to wait a few minutes for solutions more sensitive to the requirements of their domain.

The last logical step after comparing a pair of words for phonetic similarity is to aggregate the measures over a set of words. The choice of techniques depends heavily on the goal of the research. In proving historical relatedness between languages, the best approach is to sum the distance metrics across all word pairs, then scramble the words many times to see how often randomly paired words would have such a low phonetic distance between them (Kessler & Lehtonen, in press). In dialectometry, on the other hand, one normally is concerned with averaging the distance between all words for each pair of dialects. Often visualization techniques are used to help one perceive interesting patterns in the dialectological landscape. Heeringa (2004), for example, used cluster analysis and multidimensional scaling (Kruskal & Wish 1978) to analyze the relationships between the Dutch dialects and portray them on beautiful colour maps. Such summarization techniques are not specific to phonetic distance measures and so will not be discussed here. Many good ideas for portraying the relations between many different language forms can be found in general works such as those of Séguy (1973) and Goebl (1984), even though they were primarily interested in lexical relationships rather than phonological ones.

4. Prospects

It would be premature to call phonetic comparison a mature discipline, but at least it is fast becoming a discipline. More and more commonly, papers now provide some sort of quantitative assessment of their results, and in some subfields informal test suites are emerging, such as the set of cognate alignments developed by Covington (1997). Comparative assessment is also increasingly common (e.g. Kondrak 2002; Zobel & Dart 1996), due in part to the fact that many tools have been implemented as

computer programs and are being freely shared among the research community. Often it is even more useful to systematically vary techniques one at a time within one's own system and compare their efficacy. Hopefully, assessments will become more common and will be accompanied in future by significance tests, so that readers know whether increases in magnitude are actually greater than what is to be expected given random variation within the data set.

Closely related to assessment is the issue of parameter setting and training. In historical applications, most models emerge full grown from their creators' heads. However, researchers are increasingly willing to derive such numbers from empirical evidence, as did Vieregge, Rietveld & Jansen (1984). A next step may involve training: automatically tuning the parameters so that the system performs optimally, perhaps using such techniques as genetic programming (Banzhaf, Nordin, Keller & Francone 1998). The main challenge is creating a training suite big and varied enough to ensure that it is representative, lest the parameter settings inadvertently get tailored to idiosyncratic properties of the training suite.

Another trend to watch is the adaptation of algorithms used in other fields. String and sequence comparison algorithms are a favourite topic in informatics, computer science, and speech technology (Cole *et al.* 1997; Jurafsky & Martin 2000; Sankoff & Kruskal 1999). In biology as well, many of the issues involved in molecular sequencing, such as comparing DNA strands (Gusfield 1997), are much the same as in comparing words. It appears likely that the next great advance in phonetic string comparison will be inspired at least in part by advances in other fields, as linguists gain interest in these methodologies and as biologists and others seek to apply their methodological techniques to linguistic problems. An interesting array of techniques have seen some sporadic use — e.g. cross entropy by Juola (1998), n-grams by Zobel and Dart (1996) — but if fields such as speech recognition are any guide, finite state automata such as hidden Markov models may see increased application to historical linguistics (Mackay & Kondrak 2004). Whether they will supersede such techniques as Levenshtein distance is not a foregone conclusion,

however. Some of their greatest advantages, such as their speed and the ease with which they can be trained by processing huge data sets of examples, may not impress historical linguists, for whom speed may not be of the essence and huge data sets are not forthcoming.

A final advance to look forward to is a closer matching of the computational algorithm of phonetic comparison to the linguistic process it is meant to be modelling. Beginning with Grimes & Agard (1959), readers have been asked to accept, with little or no justification, that a particular phonetic measurement is a reliable index of historical distance. But if we wish to draw conclusions about the history underlying a state of affairs, we need to model as closely as possible the relevant historical events. Heggarty (2000) effectively counted a sound change that affects ten words ten times as heavily as one that affects one word, which is entirely reasonable for phenetic analysis, but we would not want to conclude there are ten times as many independent historical events. Further, phonetic distance may not be the best way to measure individual innovations. Apocope — the loss of all features in one or more segments — is dramatic phonetically, but that does not make it more of an historical event than, say, a voicing that changes just one feature.

Thus there are reasons to doubt that phonetic comparison as currently conceived is a very convincing basis for drawing strong conclusions about the cladistics of a language family. And I have not even mentioned the very real problem of parallel innovation: a sound change that has a high likelihood of occurring is weak evidence for subgrouping and should be discounted in any quantificational model of linguistic divergence. Phonetic techniques for measuring historical distance may well evolve in the direction of probabilistic finite state automata (see Raman, Newman & Patrick 1997 for an application to Chinese) or similar chaining models that take sound change events as their units of interest. Granted, computing the requisite probabilities may require more historical data than are currently available. But some progress may be made through coordinated collection of databases of diachronic information, and beyond that, one can always hope for a fortunate surprise.

Psychology Department
Washington University in St. Louis
Campus Box 1125
One Brookings Drive
St. Louis MO 63130–4899
USA
Email: bkessler@wustl.edu

References

Almeida, Antonio & Braun, Angelika, 1986. '"Richtig" und "Falsch" in phonetischer Transkription: Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten', *ZDL* 53, 158–172.

Altman, Gabriel & Lehfeldt, Werner, 1980. *Einführung in die quantitative Phonologie*, Bochum: Brockmeyer.

Austin, William M., 1957. 'Criteria for phonetic similarity', *Language* 33, 539–544.

Babitch, Rose M. & Lebrun, Eric, 1989. 'Dialectometry as computerized agglomerative hierarchical classification analysis', *JEL* 22, 83–90.

Banzhaf, Wolfgang, Nordin, Peter, Keller, Robert E. & Francone, Frank D., 1998. *Genetic Programming: an Introduction*, San Francisco, CA: Kaufmann.

Baxter, William H. & Manaster Ramer, Alexis, 2000. 'Beyond lumping and splitting: probabilistic issues in historical linguistics', in Renfrew, McMahon & Trask, 167–188.

Bernstein, Jared, Barbier, Isabella, Rosenfeld, Elizabeth & De Jong, John, 2004. 'Development and validation of an automatic spoken Spanish test', paper presented at the InSTIL/ICALL symposium NLP and Speech Technologies in Advanced Language Learning Systems, June, Venice.

Cole, Ronald (ed.), 1997. *Survey of the State of the Art in Human Language Technology*, Cambridge: C.U.P.

Connolly, John H., 1997. 'Quantifying target-realization differences', *CL&P* 11, 267–298.

Covington, Michael A., 1997. 'An algorithm to align words for historical comparison', *CL* 22, 481–496.

Dolgopolsky, Aaron B., 1986. 'A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia', in Vitalij V. Shevoroshkin and T. L. Markey (eds. and trs.), *Typology, Relationship, and Time: a Collection of Papers on Language Change and Relationship by Soviet Linguists*, Ann Arbor, MI: Karoma.

Embleton, Sheila, 2000. 'Lexicostatistics/Glottochronology: from Swadesh to Sankoff to Starotsin to future horizons', in Renfrew, McMahon & Trask, 143–165.

Fujii, Atsushi & Ishikawa, Tetsuya, 2004. 'Cross-language IR at University of Tsukuba: automatic transliteration for Japanese, English, and Korean'. *Working Notes of the Fourth NTCIR Workshop Meeting, June*, Tokyo: National Institute of Informatics.

Furui, Sadaoki, 1997. 'Speaker recognition', in Cole, 42–48.

Goebl, Hans, 1984. *Dialektometrische Studien*, Tübingen: Niemeyer.

GRIMES, JOSEPH E., 1964. 'Measures of linguistic divergence', in Horace G. Lunt (ed.), *PICL, Cambridge, Mass., August, 1962*, The Hague: Mouton, 44–50.

GRIMES, JOSEPH E. & AGARD, FREDERICK B., 1959. 'Linguistic divergence in Romance', *Language* 35, 598–604.

GUSFIELD, DAN, 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge: C.U.P.

HEERINGA, WILBERT J., 2004. *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. Ph.D. dissertation, Rijksuniversiteit Groningen.

HEGGARTY, PAUL, 2000. 'Quantifying change over time in phonetics', in Renfrew, McMahon & Trask, 531–562.

HIEMSTRA, DJOERD, 1998. 'Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus', in Peter-Arno Coppen, Hans van Halteren and Lisanne Teunissen (eds.), *Proceedings of the Eighth CLIN meeting, Nijmegen*, Amsterdam: Rodopi, 41–58.

HOPPENBROUWERS, COR & HOPPENBROUWERS, GEER, 1988. 'De featurefrequentie-methode en de classificatie van Nederlandse dialecten', *Tabu* 18, 51–92.

HOPPENBROUWERS, COR & HOPPENBROUWERS, GEER, 2001. *De indeling van de Nederlandse streektalen*, Assen: Van Gorcum.

JOOS, MARTIN, 1950. 'Description of language design', *JASA* 22, 701–708.

JUOLA, PATRICK, 1996. 'Isolated-word confusion metrics and the PGPfone alphabet', in Kemal Oflazer and Harold Somers (eds.), *NeMLaP-2: Proceedings of the Second International Conference on New Methods in Language Processing, Bilkent University*, Ankara.

JUOLA, PATRICK, 1998. 'Cross-entropy and linguistic typology', in D. M. W. Powers (ed.), *Proceedings of NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, Sydney*, 141–149.

JURAFSKY, DANIEL & MARTIN, JAMES H., 2000. *Speech and Language Processing*, Upper Saddle River, NJ: Prentice Hall.

KESSLER, BRETT, 1995. 'Computational dialectology in Irish Gaelic', in *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin*, 60–66.

KESSLER, BRETT & LEHTONEN, ANNUKKA, (in press). 'Multilateral comparison and significance testing of the Indo-Uralic question', in James Clackson, Peter Forster, and Colin Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages*. Cambridge: McDonald Inst. for Archaeological Research.

KONDRAK, GRZEGORZ, 2002. *Algorithms for Language Reconstruction*. Ph.D. dissertation, University of Toronto.

KONDRAK, GRZEGORZ & DORR, BONNIE, 2004. 'Identification of confusable drug names: a new approach and evaluation methodology'. *Proceedings of the Twentieth International Conference on Computational Linguistics, Geneva*, 952–958.

KONDRAK, GRZEGORZ, MARCU, DANIEL & KNIGHT, KEVIN, 2003. 'Cognates can improve statistical translation models', in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), Edmonton*, companion volume, 46–48.

KRUSKAL, JOSEPH B., 1999. 'An overview of sequence comparison', in David Sankoff and Joseph Kruskal (eds.), *Time Warps, String Edits, and Macromolecules*, Stanford, CA: CSLI, 1–44.

KRUSKAL, JOSEPH B. & WISH, MYRON, 1978. *Multidimensional Scaling*, Beverly Hills, CA: Sage.

LADEFOGED, PETER, 1995. *A Course in Phonetics*, New York: Harcourt Brace Jovanovich.

LAMBERT, BRUCE L., CHANG, KEN-YU & LIN, SWU-JANE, 2001. 'Effect of orthographic and phonological similarity on false recognition of drug names', *Social Science & Medicine* 52, 1843–1857.

LEVENSHTEIN, V. I., 1966. 'Binary codes capable of correcting deletions, insertions and reversals', Soviet Physics-Doklandy, 10(8), 707–710.

MACKAY, WESLEY & KONDRAK, GRZEGORZ, 2004. 'Computing word similarity with pair hidden Markov models', http://www.cs.ualberta.ca/~kondrak/WSPHMM.pdf.

MATISOFF, JAMES A., 1990. 'On megalocomparison', *Language* 66, 106–120.

MELAMED, I. DAN, 1999. 'Bitext maps and alignment via pattern recognition', *CL* 25, 107–130.

MILLER, GEORGE A. & NICELY, PATRICIA E., 1955. 'An analysis of perceptual confusions among some English consonants', *JASA* 27, 338–352.

MUTHUSAMY, YESHWANT K. & SPITZ, A. LAWRENCE, 1997. 'Automatic language identification', in Cole, 314–317.

OAKES, MICHAEL P., 2000. 'Computer estimation of vocabulary in protolanguage from word lists in four daughter languages', *JQL* 7, 233–243.

OSWALT, ROBERT L., 1970. 'The detection of remote linguistic relationships', *Computer Studies* 3, 117–129.

OSWALT, ROBERT L., 1998. 'A probabilistic evaluation of North Eurasiatic Nostratic', in J. C. Salmons and B. D. Joseph (eds.), *Nostratic: Sifting the Evidence*, Amsterdam: Benjamins, 199–216.

PAUL, HERMANN, 1880. *Prinzipien der Sprachgeschichte*, Tübingen: Niemeyer.

PIKE, KENNETH L., 1943. *Phonetics*, Ann Arbor: Univ. of Michigan Press.

RAMAN, ANAND, NEWMAN, JOHN & PATRICK, JON, 1997. 'A complexity measure for diachronic Chinese phonology', *Proceedings of the SIGPHON97 Workshop on Computational Linguistics at the ACL'97/EACL'97, Madrid.*

RENFREW, COLIN, MCMAHON, APRIL & TRASK, LARRY (eds.), 2000. *Time Depth in Historical Linguistics*, Cambridge: McDonald Inst. for Archaeological Research.

SANKOFF, DAVID & KRUSKAL, JOSEPH (eds.), 1999. *Time Warps, String Edits, and Macromolecules*, Stanford, CA: CSLI.

SÉGUY, JEAN, 1971. 'La relation entre la distance spatiale et la distance lexicale', *RLiR* 35, 335–357.

SÉGUY, JEAN, 1973. 'La dialectométrie dans *l'Atlas linguistique de la Gascogne'*, *RLiR* 37, 1–24.

SINGH, SADANAND & WOODS, DAVID R., 1971. 'Perceptual structure of 12 American English vowels', *JASA* 49, 1861–1866.

SINGH, SADANAND, WOODS, DAVID R. & BECKER, GORDON M., 1972. 'Perceptual structure of 22 prevocalic English consonants', *JASA* 52, 1698–1713.

SOMERS, HAROLD L., 1998. 'Similarity metrics for aligning children's articulation data', in *Proceedings of COLING-ACL'98, Montreal*, 1227–1232.

SOMERS, HAROLD L., 1999. 'Aligning phonetic segments for children's articulation assessment', *CL* 25, 267–275.

TOUTANOVA, KRISTINA & MOORE, ROBERT C., 2002. 'Pronunciation modeling for improved spelling correction', *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, July*, 144–151.

VIEREGGE, W. H., RIETVELD, A. C. M. & JANSEN, C. I. E., 1984. 'A distinctive feature based system for the evaluation of segmental transcription in Dutch', in M. P. R.

Van den Broecke and A. Cohen (eds.), *Proceedings of the 10th International Congress of Phonetic Sciences*, Dordrecht: Foris, 654–659.

ZOBEL, JUSTIN & DART, PHILIP, 1996. 'Phonetic string matching: lessons from informational retrieval', in Hans-Peter Frei, Donna Harman, Peter Schäuble, Ross Wilkinson (eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, Zurich*, 166–172.