Measuring Language Divergence by Intra-Lexical Comparison

T. Mark Ellison
Informatics
University of Edinburgh
mark@markellison.net

Simon Kirby
Language Evolution and Computation Research Unit
Philosophy, Psychology and Language Sciences,
University of Edinburgh
simon@ling.ed.ac.uk

Abstract

This paper presents a method for building genetic language taxonomies based on a new approach to comparing lexical forms. Instead of comparing forms cross-linguistically, a matrix of language-internal similarities between forms is calculated. These matrices are then compared to give distances between languages. We argue that this coheres better with current thinking in linguistics and psycholinguistics. An implementation of this approach, called PHILOLOGICON, is described, along with its application to Dyen et al.'s (1992) ninety-five wordlists from Indo-European languages.

1 Introduction

Recently, there has been burgeoning interest in the computational construction of genetic language taxonomies (Dyen et al., 1992; Nerbonne and Heeringa, 1997; Kondrak, 2002; Ringe et al., 2002; Benedetto et al., 2002; McMahon and McMahon, 2003; Gray and Atkinson, 2003; Nakleh et al., 2005).

One common approach to building language taxonomies is to ascribe language-language distances, and then use a generic algorithm to construct a tree which explains these distances as much as possible. Two questions arise with this approach. The first asks what aspects of languages are important in measuring inter-language distance. The second asks how to measure distance given these aspects.

A more traditional approach to building language taxonomies (Dyen et al., 1992) answers these questions in terms of **cognates**. A word in language A is said to be cognate with word in language B if the forms shared a common ancestor in the parent language of A and B. In the cognate-counting method, inter-language distance depends on the lexical forms of the languages. The distance between two languages is a function of the number or fraction of these forms which are cognate between the two languages[1]. This approach to building language taxonomies is hard to implement in toto because constructing ancestor forms is not easily automatable.

More recent approaches, such as Kondrak's (2002) and Heggarty et al's (2005) work on dialect comparison, take the synchronic word forms themselves as the language aspect to be compared. Variations on edit distance (see Kessler (2005) for a survey) are then used to evaluate differences between languages for each word, and these differences are aggregated to give a distance between languages or dialects as a whole. This approach is largely automatable, although some methods do require human intervention.

In this paper, we present novel answers to the two questions. The features of language we will compare are not sets of words or phonological forms. Instead we compare the similarities between forms, expressed as confusion probabilities. The distribution of confusion probabilities in one language is called a **lexical metric**. Section 2 presents the definition of lexical metrics and some arguments for their being good language representatives for the purposes of comparison.

The distance between two languages is the divergence their lexical metrics. In section 3, we detail two methods for measuring this divergence:

[1] McMahon and McMahon (2003) for an account of tree-inference from the cognate percentages in the Dyen et al. (1992) data.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 273–280,
Sydney, July 2006.

Kullback-Liebler (herafter KL) divergence and Rao distance. The subsequent section (4) describes the application of our approach to automatically constructing a taxonomy of Indo-European languages from Dyen et al. (1992) data.

Section 5 suggests how lexical metrics can help identify cognates. The final section (6) presents our conclusions, and discusses possible future directions for this work.

Versions of the software and data files described in the paper will be made available to coincide with its publication.

2 Lexical Metric

The first question posed by the distance-based approach to genetic language taxonomy is: what should we compare?

In some approaches (Kondrak, 2002; McMahon et al., 2005; Heggarty et al., 2005; Nerbonne and Heeringa, 1997), the answer to this question is that we should compare the phonetic or phonological realisations of a particular set of meanings across the range of languages being studied. There are a number of problems with using lexical forms in this way.

Firstly, in order to compare forms from different languages, we need to embed them in common phonetic space. This phonetic space provides granularity, marking two phones as identical or distinct, and where there is a graded measure of phonetic distinction it measures this.

There is growing doubt in the field of phonology and phonetics about the meaningfulness of assuming of a common phonetic space. Port and Leary (2005) argue convincingly that this assumption, while having played a fundamental role in much recent linguistic theorising, is nevertheless unfounded. The degree of difference between sounds, and consequently, the degree of phonetic difference between words can only be ascertained within the context of a single language.

It may be argued that a common phonetic space can be found in either acoustics or degrees of freedom in the speech articulators. Language-specific categorisation of sound, however, often restructures this space, sometimes with distinct sounds being treated as homophones. One example of this is the realisation of orthographic **rr** in European Portuguese: it is indifferently realised with an apical or a uvular trill, different sounds made at distinct points of articulation.

If there is no language-independent, common phonetic space with an equally common similarity measure, there can be no principled approach to comparing forms in one language with those of another.

In contrast, language-specific word-similarity is well-founded. A number of psycholinguistic models of spoken word recognition (Luce et al., 1990) are based on the idea of lexical neighbourhoods. When a word is accessed during processing, the other words that are phonemically or orthographically similar are also activated. This effect can be detected using experimental paradigms such as priming.

Our approach, therefore, is to abandon the cross-linguistic comparison of phonetic realisations, in favour of language-internal comparison of forms. (See also work by Shillcock et al. (2001) and Tamariz (2005)).

2.1 Confusion probabilities

One psychologically well-grounded way of describing the similarity of words is in terms of their **confusion probabilities**. Two words have high confusion probability if it is likely that one word could be produced or understood when the other was intended. This type of confusion can be measured experimentally by giving subjects words in noisy environments and measuring what they apprehend.

A less pathological way in which confusion probability is realised is in coactivation. If a person hears a word, then they more easily and more quickly recognise similar words. This coactivation occurs because the phonological realisation of words is not completely separate in the mind. Instead, realisations are interdependent with realisations of similar words.

We propose that confusion probabilities are ideal information to constitute the lexical metric. They are language-specific, psychologically grounded, can be determined by experiment, and integrate with existing psycholinguistic models of word recognition.

2.2 NAM and beyond

Unfortunately, experimentally determined confusion probabilities for a large number of languages are not available. Fortunately, models of spoken word recognition allow us to predict these probabilities from easily-computable measures of word similarity.

For example, the **neighbourhood activation model** (**NAM**) (Luce et al., 1990; Luce and Pisoni, 1998) predicts confusion probabilities from the relative frequency of words in the neighbourhood of the target. Words are *in the neighbourhood* of the target if their Levenstein (1965) edit distance from the target is one. The more frequent the word is, the greater its likelihood of replacing the target.

Bailey and Hahn (2001) argue, however, that the all-or-nothing nature of the lexical neighbourhood is insufficient. Instead word similarity is the complex function of frequency and phonetic similarity shown in equation (1). Here A, B, C and D are constants of the model, u and v are words, and d is a phonetic similarity model.

$$s = (AF(u)^2 + BF(u) + C)e^{-D.d(u,v)} \quad (1)$$

We have adapted this model slightly, in line with NAM, taking the similarity s to be the probability of confusing stimulus v with form u. Also, as our data usually offers no frequency information, we have adopted the maximum entropy assumption, namely, that all relative frequencies are equal. Consequently, the probability of confusion of two words depends solely on their similarity distance. While this assumption degrades the psychological reality of the model, it does not render it useless, as the similarity measure continues to provide important distinctions in neighbourhood confusability.

We also assume for simplicity, that the constant D has the value 1.

With these simplifications, equation (2) shows the probability of apprehending word w, out of a set W of possible alternatives, given a stimulus word w_s.

$$P(w|w_s) = e^{-d(w,w_s)}/N(w_s) \quad (2)$$

The normalising constant $N(s)$ is the sum of the non-normalised values for $e^{-d(w,w_s)}$ for all words w.

$$N(w_s) = \sum_{w \in W} e^{-d(u,v)}$$

2.3 Scaled edit distances

Kidd and Watson (1992) have shown that discriminability of frequency and of duration of tones in a tone sequence depends on its length as a proportion of the length of the sequence. Kapatsinski (2006) uses this, with other evidence, to argue that word recognition edit distances must be scaled by word-length.

There are other reasons for coming to the same conclusion. The simple Levenstein distance exaggerates the disparity between long words in comparison with short words. A word of consisting of 10 symbols, purely by virtue of its length, will on average be marked as more different from other words than a word of length two. For example, Levenstein distance between **interested** and **rest** is six, the same as the distance between **rest** and **by**, even though the latter two have nothing in common. As a consequence, close phonetic transcriptions, which by their very nature are likely to involve more symbols per word, will result in larger edit distances than broad phonemic transcriptions of the same data.

To alleviate this problem, we define a new edit distance function d_2 which scales Levenstein distances by the average length of the words being compared (see equation 3). Now the distance between **interested** and **rest** is 0.86, while that between **rest** and **by** is 2.0, reflecting the greater relative difference in the second pair.

$$d_2(w_2, w_1) = \frac{2d(w_2, w_1)}{|w_1| + |w_2|} \quad (3)$$

Note that by scaling the raw edit distance with the average lengths of the words, we are preserving the symmetric property of the distance measure.

There are other methods of comparing strings, for example string kernels (Shawe-Taylor and Cristianini, 2004), but using Levenstein distance keeps us coherent with the psycholinguistic accounts of word similarity.

2.4 Lexical Metric

Bringing this all together, we can define the lexical metric.

A lexicon L is a mapping from a set of meanings M, such as "DOG", "TO RUN", "GREEN", etc., onto a set F of forms such as /pies/, /biec/, /zielony/.

The confusion probability P of m_1 for m_2 in lexical L is the normalised negative exponential of the scaled edit-distance of the corresponding forms. It is worth noting that when frequencies are assumed to follow the maximum entropy distribution, this connection between confusion probabilities and distances (see equation 4) is the same as that proposed by Shepard (1987).

$$P(m_1|m_2;L) = \frac{e^{-d_2(L(m_1),L(m_2))}}{N(m_2;L)} \quad (4)$$

A lexical metric of L is the mapping $LM(L) : M^2 \rightarrow [0,1]$ which assigns to each pair of meanings m_1, m_2 the probability of confusing m_1 for m_2, scaled by the frequency of m_2.

$$\begin{aligned} & LM(L)(m_1,m_2) \\ = \; & P(L(m_1)|L(m_2))P(m_2) \\ = \; & \frac{e^{-d_2(L(m_1),L(m_2))}}{N(m_2;L)|M|} \end{aligned}$$

where $N(m_2;L)$ is the normalising function defined in equation (5).

$$N(m_2;L) = \sum_{m \in M} e^{-d_2(L(m),L(m_2))} \quad (5)$$

Table 1 shows a minimal lexicon consisting only of the numbers one to five, and a corresponding lexical metric. The values in the lexical metric are

	one	two	three	four	five
one	0.102	0.027	0.023	0.024	0.024
two	0.028	0.107	0.024	0.026	0.015
three	0.024	0.024	0.107	0.023	0.023
four	0.025	0.025	0.022	0.104	0.023
five	0.026	0.015	0.023	0.025	0.111

Table 1: A lexical metric on a mini-lexicon consisting of the numbers one to five.

inferred word confusion probabilities. The matrix is normalised so that the sum of each row is 0.2, ie. one-fifth for each of the five words, so the total of the matrix is one. Note that the diagonal values vary because the off-diagonal values in each row vary, and consequently, so does the normalisation for the row.

3 Language-Language Distance

In the previous section, we introduced the lexical metric as the key measurable for comparing languages. Since lexical metrics are probability distributions, comparison of metrics means measuring the difference between probability distributions. To do this, we use two measures: the symmetric Kullback-Liebler divergence (Jeffreys, 1946) and the Rao distance (Rao, 1949; Atkinson and Mitchell, 1981; Micchelli and Noakes, 2005) based on Fisher Information (Fisher, 1959). These can be defined in terms the **geometric path** from one distribution to another.

3.1 Geometric paths

The geometric path between two distributions P and Q is a conditional distribution R with a continuous parameter α such that at $\alpha = 0$, the distribution is P, and at $\alpha = 1$ it is Q. This conditional distribution is called the **geometric** because it consists of normalised weighted geometric means of the two defining distributions (equation 6).

$$R(\bar{w}|\alpha) = P(\bar{w})^{\alpha} Q(\bar{w})^{1-\alpha} / k(\alpha;P,Q) \quad (6)$$

The function $k(\alpha;P,Q)$ is a normaliser for the conditional distribution, being the sum of the weighted geometric means of values from P and Q (equation 7). This value is known as the Chernoff coefficient or Helliger path (Basseville, 1989). For brevity, the P, Q arguments to k will be treated as implicit and not expressed in equations.

$$k(\alpha) = \sum_{\bar{w} \in W^2} P(\bar{w})^{1-\alpha} Q(\bar{w})^{\alpha} \quad (7)$$

3.2 Kullback-Liebler distance

The first-order (equation 8) differential of the normaliser with regard to α is of particular interest.

$$k'(\alpha) = \sum_{\bar{w} \in W^2} \log \frac{Q(\bar{w})}{P(\bar{w})} P(\bar{w})^{1-\alpha} Q(\bar{w})^{\alpha} \quad (8)$$

At $\alpha = 0$, this value is the negative of the Kullback-Liebler distance $KL(P|Q)$ of Q with regard to P (Basseville, 1989). At $\alpha = 1$, it is the Kullback-Liebler distance $KL(Q|P)$ of P with regard to Q. Jeffreys' (1946) measure is a symmetrisation of KL distance, by averaging the commutations (equations 9,10).

$$\begin{aligned} KL(P,Q) &= \frac{KL(Q|P) + KL(P|Q)}{2} \quad (9) \\ &= \frac{k'(1) - k'(0)}{2} \quad (10) \end{aligned}$$

3.3 Rao distance

Rao distance depends on the second-order (equation 11) differential of the normaliser with regard to α.

$$k''(\alpha) = \sum_{\bar{w} \in W^2} \log^2 \frac{Q(\bar{w})}{P(\bar{w})} P(\bar{w})^{1-\alpha} Q(\bar{w})^{\alpha} \quad (11)$$

Fisher information is defined as in equation (12).

$$FI(P,x) = -\int \frac{\partial^2 \log P(y|x)}{\partial x^2} P(y|x) dy \quad (12)$$

Equation (13) expresses Fisher information along the path R from P to Q at point α using k and its first two derivatives.

$$\mathrm{FI}(R,\alpha) = \frac{k(\alpha)k''(\alpha) - k'(\alpha)^2}{k(\alpha)^2} \qquad (13)$$

The Rao distance $r(P,Q)$ along R can be approximated by the square root of the Fisher information at the path's midpoint $\alpha = 0.5$.

$$\mathrm{r}(P,Q) = \sqrt{\frac{k(0.5)k''(0.5) - k'(0.5)^2}{k(0.5)^2}} \qquad (14)$$

3.4 The PHILOLOGICON algorithm

Bringing these pieces together, the PHILOLOGICON algorithm for measuring the divergence between two languages has the following steps:

1. determine their joint confusion probability matrices, P and Q,
2. substitute these into equation (7), equation (8) and equation (11) to calculate $k(0)$, $k(0.5)$, $k(1)$, $k'(0.5)$, and $k''(0.5)$,
3. and put these into equation (10) and equation (14) to calculate the KL and Rao distances between between the languages.

4 Indo-European

The ideal data for reconstructing Indo-European would be an accurate phonemic transcription of words used to express specifically defined meanings. Sadly, this kind of data is not readily available. However, as a stop-gap measure, we can adopt the data that Dyen et al. collected to construct a Indo-European taxonomy using the cognate method.

4.1 Dyen et al's data

Dyen et al. (1992) collected 95 data sets, each pairing a meaning from a Swadesh (1952)-like 200-word list with its expression in the corresponding language. The compilers annotated with data with cognacy relations, as part of their own taxonomic analysis of Indo-European.

There are problems with using Dyen's data for the purposes of the current paper. Firstly, the word forms collected are not phonetic, phonological or even full orthographic representations. As the authors state, the forms are expressed in sufficient detail to allow an interested reader acquainted with the language in question to identify which word is being expressed.

Secondly, many meanings offer alternative forms, presumably corresponding to synonyms. For a human analyst using the cognate approach, this means that a language can participate in two (or more) word-derivation systems. In preparing this data for processing, we have consistently chosen the first of any alternatives.

A further difficulty lies in the fact that many languages are not represented by the full 200 meanings. Consequently, in comparing lexical metrics from two data sets, we frequently need to restrict the metrics to only those meanings expressed in both the sets. This means that the KL divergence or the Rao distance between two languages were measured on lexical metrics cropped and rescaled to the meanings common to both data-sets. In most cases, this was still more than 190 words.

Despite these mismatches between Dyen et al.'s data and our needs, it provides an testbed for the PHILOLOGICON algorithm. Our reasoning being, that if successful with this data, the method is reasonably reliable. Data was extracted to language-specific files, and preprocessed to clean up problems such as those described above. An additional data-set was added with random data to act as an outlier to root the tree.

4.2 Processing the data

PHILOLOGICON software was then used to calculate the lexical metrics corresponding to the individual data files and to measure KL divergences and Rao distances between them. The program NEIGHBOR from the PHYLIP[2] package was used to construct trees from the results.

4.3 The results

The tree based on Rao distances is shown in figure 1. The discussion follows this tree except in those few cases mentioning differences in the KL tree.

The standard against which we measure the success of our trees is the conservative traditional taxonomy to be found in the Ethnologue (Grimes and Grimes, 2000). The fit with this taxonomy was so good that we have labelled the major branches with their traditional names: Celtic, Germanic, etc. In fact, in most cases, the branch-internal divisions — eg. Brythonic/Goidelic in Celtic, Western/Eastern/Southern in Slavic, or

[2] See **http://evolution.genetics.washington.edu/phylip.html**.

Western/Northern in Germanic — also accord. Note that PHILOLOGICON even groups Baltic and Slavic together into a super-branch Balto-Slavic.

Where languages are clearly out of place in comparison to the traditional taxonomy, these are highlighted: visually in the tree, and verbally in the following text. In almost every case, there are obvious contact phenomena which explain the deviation from the standard taxonomy.

Armenian was grouped with the Indo-Iranian languages. Interestingly, Armenian was at first thought to be an Iranian language, as it shares much vocabulary with these languages. The common vocabulary is now thought to be the result of borrowing, rather than common genetic origin. In the KL tree, Armenian is placed outside of the Indo-Iranian languages, except for Gypsy. On the other hand, in this tree, Ossetic is placed as an outlier of the Indian group, while its traditional classification (and the Rao distance tree) puts it among the Iranian languages. Gypsy is an Indian language, related to Hindi. It has, however, been surrounded by European languages for some centuries. The effects of this influence is the likely cause for it being classified as an outlier in the Indo-Iranian family. A similar situation exists for Slavic: one of the two lists that Dyen et al. offer for Slovenian is classed as an outlier in Slavic, rather than classifying it with the Southern Slavic languages. The other Slovenian list is classified correctly with Serbocroatian. It is possible that the significant impact of Italian on Slovenian has made it an outlier. In Germanic, it is English that is the outlier. This may be due to the impact of the English creole, Takitaki, on the hierarchy. This language is closest to English, but is very distinct from the rest of the Germanic languages. Another misclassification also is the result of contact phenomena. According to the Ethnologue, Sardinian is Southern Romance, a separate branch from Italian or from Spanish. However, its constant contact with Italian has influenced the language such that it is classified here with Italian. We can offer no explanation for why Wakhi ends up an outlier to all the groups.

In conclusion, despite the noisy state of Dyen et al.'s data (for our purposes), the PHILOLOGICON generates a taxonomy close to that constructed using the traditional methods of historical linguistics. Where it deviates, the deviation usually points to identifiable contact between languages.

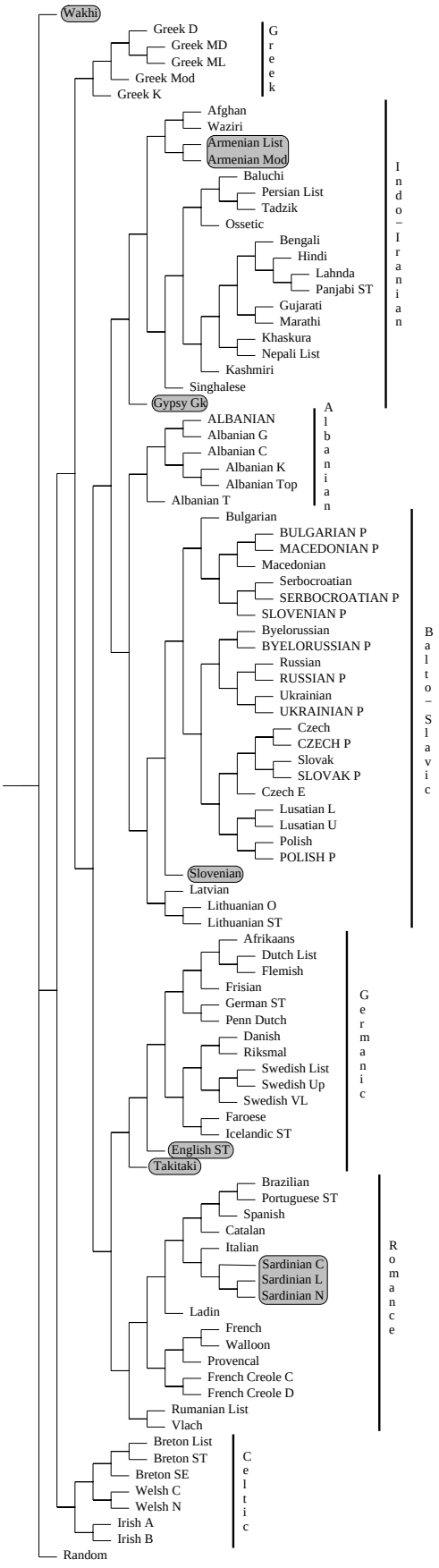

Figure 1: Taxonomy of 95 Indo-European data sets and artificial outlier using PHILOLOGICON and PHYLIP

5 Reconstruction and Cognacy

Subsection 3.1 described the construction of geometric paths from one lexical metric to another. This section describes how the synthetic lexical metric at the midpoint of the path can indicate which words are cognate between the two languages.

The synthetic lexical metric (equation 15) applies the formula for the geometric path equation (6) to the lexical metrics equation (5) of the languages being compared, at the midpoint $\alpha = 0.5$.

$$R_{\frac{1}{2}}(m_1, m_2) = \frac{\sqrt{P(m_1|m_2)Q(m_1|m_2)}}{|M|k(\frac{1}{2})} \quad (15)$$

If the words for m_1 and m_2 in both languages have common origins in a parent language, then it is reasonable to expect that their confusion probabilities in both languages will be similar. Of course different cognate pairs m_1, m_2 will have differing values for R, but the confusion probabilities in P and Q will be similar, and consequently, the reinforce the variance.

If either m_1 or m_2, or both, is non-cognate, that is, has been replaced by another arbitrary form at some point in the history of either language, then the P and Q for this pair will take independently varying values. Consequently, the geometric mean of these values is likely to take a value more closely bound to the average, than in the purely cognate case.

Thus rows in the lexical metric with wider dynamic ranges are likely to correspond to cognate words. Rows corresponding to non-cognates are likely to have smaller dynamic ranges. The dynamic range can be measured by taking the Shannon information of the probabilities in the row.

Table 2 shows the most low- and high-information rows from English and Swedish (Dyen et al's (1992) data). At the extremes of low and high information, the words are invariably cognate and non-cognate. Between these extremes, the division is not so clear cut, due to chance effects in the data.

English	**Swedish**	$10^4(h - \bar{h})$
Low Information		
we	vi	-1.30
here	her	-1.19
to sit	sitta	-1.14
to flow	flyta	-1.04
wide	vid	-0.97
	:	
scratch	klosa	0.78
dirty	smutsig	0.79
left (hand)	vanster	0.84
because	emedan	0.89
High Information		

Table 2: Shannon information of confusion distributions in the reconstruction of English and Swedish. Information levels are shown translated so that the average is zero.

6 Conclusions and Future Directions

In this paper, we have presented a distance-based method, called PHILOLOGICON, that constructs genetic trees on the basis of lexica from each language. The method only compares words language-internally, where comparison seems both psychologically real and reliable, and never cross-linguistically, where comparison is less well-founded. It uses measures founded in information theory to compare the intra-lexical differences.

The method successfully, if not perfectly, recreated the phylogenetic tree of Indo-European languages on the basis of noisy data. In further work, we plan to improve both the quantity and the quality of the data. Since most of the mis-placements on the tree could be accounted for by contact phenomena, it is possible that a network-drawing, rather than tree-drawing, analysis would produce better results.

Likewise, we plan to develop the method for identifying cognates. The key improvement needed is a way to distinguish indeterminate distances in reconstructed lexical metrics from determinate but uniform ones. This may be achieved by retaining information about the distribution of the original values which were combined to form the reconstructed metric.

References

C. Atkinson and A.F.S. Mitchell. 1981. Rao's distance measure. *Sankhyā*, 4:345–365.

Todd M. Bailey and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44:568–591.

Michle Basseville. 1989. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, December.

D. Benedetto, E. Caglioti, and V. Loreto. 2002. Language trees and zipping. *Physical Review Letters*, 88.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An indo-european classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5).

R.A. Fisher. 1959. *Statistical Methods and Scientific Inference*. Oliver and Boyd, London.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426:435–439.

B.F. Grimes and J.E. Grimes, editors. 2000. *Ethnologue: Languages of the World*. SIL International, 14th edition.

Paul Heggarty, April McMahon, and Robert McMahon, 2005. *Perspectives on Variation*, chapter From phonetic similarity to dialect classification. Mouton de Gruyter.

H. Jeffreys. 1946. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A*, 186:453–461.

Vsevolod Kapatsinski. 2006. Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. Technical Report 27, Indiana University Speech Research Lab.

Brett Kessler. 2005. Phonetic comparison algorithms. *Transactions of the Philological Society*, 103(2):243–260.

Gary R. Kidd and C.S. Watson. 1992. The "proportion-of-the-total-duration rule for the discrimination of auditory patterns. *Journal of the Acoustic Society of America*, 92:3109–3118.

Grzegorz Kondrak. 2002. *Algorithms for Language Reconstruction*. Ph.D. thesis, University of Toronto.

V.I. Levenstein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Paul Luce and D. Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19:1–36.

Paul Luce, D. Pisoni, and S. Goldinger, 1990. *Cognitive Models of Speech Perception: Psycholinguistic and Computational Perspectives*, chapter Similarity neighborhoods of spoken words, pages 122–147. MIT Press, Cambridge, MA.

April McMahon and Robert McMahon. 2003. Finding families: quantitative methods in language classification. *Transactions of the Philological Society*, 101:7–55.

April McMahon, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. Swadesh sublists and the benefits of borrowing: an andean case study. *Transactions of the Philological Society*, 103(2):147–170.

Charles A. Micchelli and Lyle Noakes. 2005. Rao distances. *Journal of Multivariate Analysis*, 92(1):97–115.

Luay Nakleh, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an ie dataset. *Transactions of the Philological Society*, 103(2):171–192.

J. Nerbonne and W. Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of SIGPHON-97: 3rd Meeting of the ACL Special Interest Group in Computational Phonology*.

B. Port and A. Leary. 2005. Against formal phonology. *Language*, 81(4):927–964.

C.R. Rao. 1949. On the distance between two populations. *Sankhyā*, 9:246–248.

D. Ringe, Tandy Warnow, and A. Taylor. 2002. Indo-european and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

R.N. Shepard. 1987. Toward a universal law of generalization for physical science. *Science*, 237:1317–1323.

Richard C. Shillcock, Simon Kirby, Scott McDonald, and Chris Brew. 2001. Filled pauses and their status in the mental lexicon. In *Proceedings of the 2001 Conference of Disfluency in Spontaneous Speech*, pages 53–56.

M. Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American philosophical society*, 96(4).

Monica Tamariz. 2005. *Exploring the Adaptive Structure of the Mental Lexicon*. Ph.D. thesis, University of Edinburgh.