April McMahon and Robert McMahon. *Language Classification by the Numbers*. Oxford: Oxford University Press, 2005. xvii + 265 pages, ISBN 0-19-9270901-2, ISBN 0-19-9270902-0 (paperback); £57,50, £22,50 (paperback).

Review

John Nerbonne

Evolutionary biologists have developed powerful, largely automated techniques for inferring the common history of groups of organisms or species based on their shared characteristics, a field known as (computational) PHYLOGENETICS or CLADISTICS. These techniques are regularly applied to human genetic material to infer the genetic history of plant and animal populations, including groups of people, e.g., the peoples that settled the Pacific islands, but the techniques are abstract enough and robust enough to be applied to linguistic material as well. Naturally the application of quantitative techniques to a novel field raises a range of fundamental questions, and April and Robert McMahon have written a gentle introduction to the application of phylogenetics to questions in historical linguistics in which they address many of the fundamental questions concerning the application of quantitative and computational techniques, including phylogenetics, to questions in historical linguistics.

One can imagine the data organised in a large table of languages × characteristics, something along the following lines:

	MAN?	FISH?	…	V2	OV	…	p^h	t^h	ASP.	…
English	+	+		-	-		+	+	+	
German	+	+		+	%		+	+	+	
Dutch	+	+		+	%		-	-	-	
French	-	+		-	-		-	-	-	
Irish	-	+		-	-		+	+	+	

The characteristics in the first two columns are intended to refer to the existence of cognates so frequently invoked in historical linguistics. All the languages have cognates for the word FISH, and the first three for the word MAN. But nothing in the procedures prevents one from referring to more abstract properties such as verb position (V2), relative position of verb and object (OV) or aspiration in the phoneme /p/, or aspiration in /t/, or aspiration in voiceless stops in general (marked 'ASP.'), properties of eminent typological

interest. Indeed Dunn et al. (2005) have already applied phylogenetic analysis to a database of primarily morphosyntactic information about Melanesian languages. The values in the cells of the table might be binary +/-, as most of those above, but they might be more complex, too. We used '%' to note situations in which the OV characteristic is required in some cases (here, in subordinate clauses), but not in others. Beginning from a table of properties such as the one above, phylogenetic software seeks explanations in the form of hypotheses about common ancestor languages. If two or more languages (or varieties) have descended from the same ancestor, this explains why they have come to share some properties.

Typologists will wish to follow the developments in this emerging area for two reasons. First, many cladistic techniques assume that the methods are applied to statistically independent characteristics, and the question of statistical independence is fundamentally typological, which means that typology will need to contribute to the development of this emerging specialisation, particularly as the databases of material become larger and more complex. The example above would clearly be unacceptable as input to a cladistic procedure. Once we know that a language has an aspirated phoneme /t^h/, then we also know, typologically, that if it has any fortis bilabial stop /p/, it will also be aspirated (/p^h/). In larger collections, typological constraints may not be obvious, so that additional quantitative and statistical techniques will be needed in conjunction with cladistics to address these questions, which suggests in turn that quantitative typology and quantitative historical linguistics will need to track each other's work.

The second reason is related to the first, and is more fundamental. At an abstract level typology and phylogeny---and let's add, geography in the form of areal linguistics---compete in offering explanations of why languages have their specific characteristics. If the question is why a particular English variety realizes the initial consonant in the word 'tide' ([t^haɪd]) as it does, then we look to historical phonology for an explanation of why this is a /t/, comparing (at least) the predictable counterparts in other Germanic languages (and others). If we ask more exactly why the /t/ in tide is aspirated ([t^h]), then it is important to note that aspiration is typologically a property of (the fortis variants of) stop contrasts in English, and that the aspiration of the /t/ simply follows from the aspiration properties of the stop contrasts in the language. If our database contains information about Scottish English dialects, where the fortis-lenis series of stops is sometimes realized as unaspirated vs. prevoiced, then we need in addition an appeal to areal information, the

information that the variety we are interested in is not near the area where English stops are realised as unaspirated vs. prevoiced.

Strong Points of LCN

There are many reasons to recommend LCN. There are intelligent discussions about what one can learn from lists of putatively cognate words, and in particular, if one is interested in recognising potential signals of historical relatedness among languages, what pitfalls one may encounter in the form of chance resemblance, confounds due to onomatopoeia, or mixed signals due to contact influences (borrowing).

Two aspects of the book stand out. First, McMahon & McMahon take great pains to consider the role of borrowings in historical linguistics (Sections 3.2.2, 4.2 and elsewhere). They point out the danger of confounding that borrowing implies if shared characteristics are always taken as evidence of shared phylogenetic history, and they explain the philological care needed to avoid the use of material which has been borrowed. But importantly, they plead for the use of techniques which simply recognise that some characteristics will not be inherited "from above" in linguistic genealogies, but may be inferred to have been transferred "horizontally". Clearly this does not obviate the need for work aimed at distinguishing the two processes of retaining or acquiring characteristics (p.174), and the authors are appropriately encouraging about the linguistic work which might seek to confirm or disconfirm hypotheses adduced about the sources of material.

Second, they succeed in explaining NETWORK MODELS, which operationalise the inferences needed when one admits hypotheses of borrowing, quite clearly (Chap. 6), even including a brief biological example. They acknowledge a debt to Bryant, Filimon & Gray (2005), which is indeed a more detailed presentation of the statistics and models used here, a worthwhile reference to the more technically interested reader.[1]

Clearly a central goal of LCN is to engage historical linguists and to convince them that incursions into phylogenetics are worthwhile, and to anticipate objections which may be outdated. So there are lengthy

[1] To which we add McMahon & McMahon's reference to the standard text, Felsenstein (2003). We also found the other linguistics papers in Mace, Holden & Shennan (2005) useful.

discussions of the potential advantages of using quantitative techniques in historical linguistics, the need to be cautious in inferring anything with respect to the dating of the historical developments, and, with appropriately cautious qualifications, hints at points at which quantitative techniques hold the promise of improving on traditional methods. Given that *Language* has now published a lengthy article on the use of phylogenetic techniques in historical linguistics (Nakleh et al. 2005), it is clear that the field is taking this line of work very seriously, and that a book making more of the work accessible appears at an opportune moment.

Problems with LCN

Although LCN will be useful for readers interested in the dialogue between historical linguistics and phylogenetics, it also has some problems. The title suggests that attention might be paid to classification of other sorts, such typological classification or areal classification. The latter is discussed briefly in Chap. 8, and the former is simply absent. The distinction between PHENETIC techniques, which identify most similar groups, and phylogenetic techniques, which seek to reconstruct a simplest history of shared characteristics, is made only in passing (p.158) and without explanation or illustration, even though the distinction was encountered earlier. The basic phylogenetic techniques, those which ignore the possibility of horizontal transfer, or borrowing, are explained only very briefly (p.73). Readers would have benefited from a more patient presentation and perhaps a simple example.

The greatest problem stems from LCN's attempt to present the phylogenetic work nontechnically. It is clear that the authors want to make phylogenetics understandable to historical linguists, even those with little background in statistics or computation, and this requires that their explanations be sketchy at times. The difficulty is that they do not maintain a consistent level of explanation in the book. So while the authors go to the trouble of explaining very basic notions such as *p*-values in statistical testing, they present the binomial formula in an impossibly quick paragraph (p.54), and then do not bother explaining maximum likelihood estimation (even though the concept is used to distinguish different approaches, pp.99-100, p.197), or similarly “Bayesian Markov chain Monte-Carlo simulation” (p.197). A less technically versed reader who wishes to understand the material more thoroughly is encouraged in early parts of the book, but frustrated later.

The technical presentation is correct throughout, but occasionally thin when only a bit more effort would be required. We note two examples.

- On p.171 the authors analyse a difference in percentages of Spanish borrowings into Quechua or Aymaran varieties. At issue is the difference between two lists of 30 elements each, where they find on average 2.7% borrowings in the one and 6.7% in the other, concluding that the "difference is significant at the $p<0.001$ level (paired *t*-test; $t=-4.2$, df=18)", without, however, explaining what (numerical) paired measure was applied to what pairs of items. Leafing back to p.166, we see that they are dealing with 19 varieties, leading us to guess that they are analysing the difference *per variety* in the number of borrowings involving concepts from the one list as opposed to those involving concepts from the other.

- On p.94 the authors analyse the difference in the percentage of borrowings in the first 100 words of the Swadesh list (8.6%) as opposed to those in the second half (15.7%). Then they appear to total these across five languages examined and analyse the 2×2 table via χ^2, concluding that the difference is significant "($\chi^2= 10.7$, $p<0.001$)". If they weren't totalling over the five samples, we'd never get significance, but this is the only hint we get.

There is also an inconclusive discussion about the size of data lists required (p.95), even though dialectological work they cite has suggested a means of quantifying the reliability of the material under comparison (see the use of Cronbach's α presented in Heeringa et al. 2002 and elsewhere). It is very surprising not to see any discussion of or indeed any mention of Cavalli-Sforza & Wang (1986), perhaps the *locus classicus* for discussions of quantitative perspectives on language history which derive inspiration from population genetics, and, most regrettably, no note is taken of Kondrak (2002), arguably the best work to-date on opportunities for improving historical phonology using computational techniques.

Criticism of Groningen work on Pronunciation Distance

Although McMahon & McMahon focus on classification for the purposes of historical linguistics, they include a section very critical of work that has been done in Groningen on measuring and classifying language varieties for the purpose of dialectology. Since quantitative analyses of dialect similarity, just as quantitative analyses of historical linguistics, will ultimately need to interact with typology, we discuss this as well here.

It is worth mentioning that dialectology and historical "classification" may not be exactly the same endeavour--in Groningen we have focused on developing techniques seeking to identify the signal of geographic provenance in dialect speech, while historical linguistics must try to identify signals of historical "relatedness". One important difference between these two is their relation to geography, which influences the distribution of dialectal varieties massively, but not necessarily discretely. Thus Heeringa & Nerbonne (2001) show how the dialectal analysis provides an analytical foundation for the notion "dialect continuum", in which classification into discrete groups, the very heart of phylogenetic analysis, plays no role. But we examine each of McMahon & McMahon's individual criticisms below.

The work in Groningen has experimented with techniques for measuring the difference in pronunciation between words. These techniques are varieties of the sequence distance measure known as EDIT DISTANCE or LEVENSHTEIN DISTANCE. At a rough degree of approximation, this procedure first aligns corresponding segments in the phonetic transcriptions of two words, where segments may align with the null segment,∅, corresponding to insertions and deletions. The distance between the words is then the sum of the distances between corresponding segments. The technique is fully automated and may be applied directly to the sort of material often found in dialect atlases—lists of pronunciations of the same words as they are pronounced in a large number of data collection sites.

McMahon & McMahon first criticise that Nerbonne and Heeringa's "earlier work calculated edit distance in the simplest possible way meaning that the pair [a,t] count as different to the same degree as [a, ɔ]", citing Nerbonne & Heeringa (1997:11). But in fact the paper they cite *focuses* on how to differentiate such sounds more subtly, exploiting phonetic and phonological features for this purpose. The discussion of this takes up the more than half of the brief paper, starting on p.12. Further, the issue of segmental similarity has been a focus of the work in Groningen from the very beginning (Nerbonne et al. 1996), culminating in Heeringa (2004), which devotes one 52-pp. chapter to the question of how to measure similarity of two phonetic segments using phonetic feature systems, comparing three different systems in detail: first, a

feature system borrowed from *The Sound Pattern of English*; second, the Vieregge et al. (1984) system; and third, a system developed for phonetic segment characterisation by Almeida & Braun (1986).[2]

McMahon & McMahon then acknowledge that later work in Groningen indeed does use variable segment costs, depending on the segments involved (LCN, p.211), but regret that the authors "do not illustrate these different replacement costs" (p.211), and complain as well that "it is difficult to replicate these calculations" (p.212) and that the procedure does "not always seem to be derived in principled way" (p.213). But Nerbonne & Heeringa (1997:13) uses the feature system developed by Vieregge (1984), and Nerbonne, Heeringa & Kleiweg (1999) an alternative system inspired by Chomsky and Halle's *Sound Pattern of English*, developed by Hoppenbrouwers & Hoppenbrouwers (1988), which performed less satisfactorily. While readers are explicitly referred to the papers explicating segment differences, our early research reports contain indeed only illustrations of Vieregge et al.'s or Hoppenbrouwers & Hoppenbrouwers's systems because these systems were only *used* in the reports, whose focus lay elsewhere. On the other hand, Heeringa (2004: Ch. 3-4, pp. 27-120) is completely explicit about a range of segmental distance measures. We encourage readers to compare the level of explicitness there with that in LCN, Sections 8.4-8.5.

McMahon & McMahon's criticism that one should not value a substitution's contribution to distance as the same as that of an insertion plus a deletion (p.211) would once have been *a propos*, but this practice was superseded even in the second of the papers McMahon & McMahon cite (Heeringa, 2002:446). For our current thinking on this, see Heeringa (2004:Ch.5), in which all operations—substitutions, insertions and deletions—have the same chance to contribute roughly the same to distance, but where all operations are weighted by the segment distance discussed above and realized variously—through feature differences or through acoustic differences. McMahon & McMahon are correct in criticising that we mostly ignore metatheses in our published work. We have, however, experimented with metathesis operators, and we can report that the effect of allowing metathesis to be considered is negligible with respect to correctness because metatheses are so infrequent (and that their inclusion involves a (mild) slowdown because more

[2] Available at http://dissertations.ub.rug.nl/faculties/arts/2004/w.j.heeringa/ (and individual chapters at http://www.let.rug.nl/heeringa/dialectology/) There is also a second, 42-pp. chapter on acoustically based measures, which, in general outperform the feature-based methods (but only slightly).

hypotheses need to be considered). But see Heeringa (2004:125-126) for an explanation of why the inclusion of metathesis operators is complex in systems using segment distances.

Even though McMahon & McMahon regret that the segment comparisons used by Groningen are not illustrated, they nonetheless criticise that the differences between front and back vowels are regarded as too large in comparison with the differences high and low vowels. This is indeed an area in which linguistic wisdom may play a greater role, but at least two words of caution are appropriate. First, McMahon & McMahon criticise that, in our work, "the number of height contrasts should be 4, [...] but that [...] the number of contrasts in frontness/backness should be 6, which is staggeringly high". We repeat the point made above that McMahon & McMahon might have more advantageously followed the reference provided to Vieregge et al. in order to understand why the contributions to distance are quantified as they are, but the reason has nothing to do with the number of possible contrasts, either universally or language specifically. These segment-difference metrics were proposed by Vieregge et al. (1984) and Almeida and Braun (1986) as a means of scoring the deviance of perceptions in which the one value is confused for another (and which lead to mistranscription, which they wished to quantify). The idea behind their scheme is that it is more deviant, for example, to perceive an /i/ as an /u/ than it is to perceive an /i/ as an /æ/. This is disputable, but it is a reasonable starting point for a theory like ours, in which one wants to quantify the signal of geographic provenance in speech. Very deviant pronunciations are strong signals of different provenance.

Second, and here we probably do not differ from McMahon & McMahon in our ultimate goals, but rather in our judgement of the current state of the art, we need to insist that measurement proposals be empirically validated, and that the validation be as independent as possible of the classification task. To appreciate how complex the question of segment similarity is, consider the range of 6-20 distinctive features that may be used to define segments, the possibility of weighting these, as well the possibility of refinements that may be called for to deal with diphthongs and segmental length as well as secondary articulations such as rhoticisation, labialisation, palatalisation, etc. Given the huge number of parameter combinations that may be invoked in defining distance measures, the danger looms that researchers are merely industrious, not insightful in developing techniques. A guard against this is to seek independent validation. This line of thinking has led us in Groningen to be sceptical of validating the pronunciation distance metric we have developed only on the basis of the groupings into dialect areas it ultimately leads to (through similarity

clustering, or what phylogeneticists call phenetic techniques), although we have attempted to quantify this rigorously (Heeringa et al. 2002, Heeringa et al. 2006). The problem is simply that the clustering required to establish the groups is known to be unstable, meaning that small differences in input data may lead to large differences in output results. As long as this is true, we should not be satisfied by simply examining clustered results, or for that matter, results reported from phylogenetic analysis, such as McMahon & McMahon's (see e.g. pp. 221-223); instead, we need to mistrust the results of clustering (phenetic analysis), and we need to look elsewhere for validation. We have suggested in Groningen that, since we are ultimately interested in identifying signals of geographic provenance, and since the social function of these signals is to identify the geographic provenance of interlocutors, we might use (non-linguists') judgements of deviance with respect to the judges' "home" dialects (Gooskens 2004) to validate computational results, and we have chosen to validate our measurements with respect to dialect speakers' perceptions (Heeringa & Gooskens 2003; Heeringa 2004; Heeringa et al., 2006).[3]

We concede that the validating material in the case of historical linguistics, the focus of McMahon & McMahon's interest, may be very difficult to determine, but we disagree that our dialectal work "is to a great extent uncorroborated" (p.213). On the contrary, the techniques have been applied to Dutch, German, American English, Sardinian, Norwegian and Bulgarian with a degree of success which specialists in the dialects of these languages have recognised. The contributions of various aspects of the measurement technique have been tested quantitatively in their ability to match the judgements of dialect speakers, including the sensitivity of the measure to segment order and to phonological context, the (non-)use of length normalisation, and the linguistic constraint that all alignments respect the consonant/vowel distinction (Heeringa et al. 2006).

Like Kessler (1995), we have not been successful in Groningen in demonstrating the superiority of measurements in which more sensitive segment contrasts are used (Heeringa 2004:Ch.7), but we hasten to add that we have approached the problem more rigorously than McMahon & McMahon (Section 8.4), who simply include feature-based segment differences in their measurements, and then conclude that they are successful when the resulting phylogenetic tree seems satisfactory. Their results are not different from the

[3] In fact, we have also explored a third line of validation for cases in which independent corroboration is unavailable, namely the degree to which measurements expose geographic coherence in the data (Nerbonne & Kleiweg, forthcoming).

use of more sensitive segment distances in Groningen. There, too, feature-based techniques led to perfectly acceptable measurements (Nerbonne & Heeringa 1997). But when we compare measurements with and without more sensitive segment differences, we conclude that the segment differences do not "earn" the additional complexity they introduce. Further experimentation may reverse this conclusion, but it may just be that the essential distinction in dialectology is just "same vs. different" so that it is beneficial to ignore finer phonological detail. Other explanations are also possible. For example, simpler measurement schemes may counteract too much ambition in the detail of field workers' and transcribers' records; or alternatively, the simpler schemes may neutralize tendencies of field workers and transcribers to impose, albeit subconsciously, tendencies they suspect in the data. The data is also extremely detailed, making complete systems for segment distance quite complex. On the other hand, Mackay & Kondrak (2005) and Kondrak (2006), focusing on cognate identification, demonstrate the superiority of versions of edit distance in which segment distances are automatically induced, which makes us optimistic about obtaining segment distances through automatic techniques.

In addition, we note that we are applying the measurement to independently collected, and quite complex, dialect atlas material, not merely the eleven Romance varieties in McMahon & McMahon's chapter 8.4, and the 20 varieties of Germanic in chapter 8.5. We normally apply analyses to hundreds of varieties, which are often transcribed in painstaking detail. The (publicly available) American English LAMSAS data distinguishes over 1,100 different vowels at over 450 sites, for example (different combinations of vowels symbols with additional diacritics for length, nasality, height, advancement, rounding, r-colouring, and stress). These complicate the definition of a segment distance metric considerably.

A particularly puzzling aspect of McMahon & McMahon's critique is their discussion of what they call "compatibility", which they define as "ensuring that we are comparing like with like" (p.209), and which they see threatened not only by metathesis (see above), but also by "elisions of vowels, changes of consonants to glides and subsequently to the second elements of diphthongs, lengthenings and shortenings of segments, and the introduction of new segments" (p. 212). They seem to be worried that our techniques may not compare the relevant segments within words whose pronunciation distance is being measured. But the edit distance technique we (as well as Kessler and Kondrak) have championed produces an alignment of

string pairs automatically, and, although it is not perfect, we are certain that the alignment errors are not a major problem.

McMahon & McMahon also ask "how Nerbonne et al. select which strings to compare in the first place", but this question is answered in our publications, which are based on published material from dialect atlases, which is also available to other researchers. McMahon & McMahon ask in particular what is done "when strings are not true cognates", and this question arises frequently. Whenever dialect atlases separate material illustrating pronunciation differences from material illustrating lexical differences (such as LAMSAS), comparisons have been restricted to using only the relevant material, i.e. material which consists of different pronunciations of the same lexeme. In cases where the atlas materials do not separate lexical and pronunciation material (such as the Dutch RND), we have often simply applied the pronunciation algorithm to *all* the material, mixing pronunciation and lexical differences. This had the virtue of not requiring us to select material, perhaps with an (unintentional) bias. Heeringa & Nerbonne (2006) investigates the effect of mixing these levels. The applications of the measurement to mixed material in fact correlate very highly with those applied to pure pronunciation material (r=0.99).

Finally, McMahon & McMahon criticise that we validate our results by comparing them to "the opinions of expert dialectologists", which we in fact do find very important, but, as we noted above, we have proceeded well beyond this to validations comparing our calculations to the judgements of dialect speakers. As far as we know, the Groningen group has contributed more to the validation of computations of linguistic distance more than any other group working in this area.

John Nerbonne

Centre for Language and Cognition,

P.O.Box 716

University of Groningen

9700 AS Groningen

Netherlands

j.nerbonne@rug.nl

Acknowledgments

My thanks to Wilbert Heeringa, Therese Leinonen and April McMahon for discussion of the material here.

As must be obvious, this does not mean that either agrees with all of what is said.

References

Almeida, Almerindo & Angelika Braun (1986). 'Richtig' und 'Falsch' in phonetischer Transkription: Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik* LIII-2:158-172.

Bryant, David, Flavia Filimon, & Russell D. Gray (2005). Untangling our past: Languages, trees, splits and networks. In Mace, Holden & Shennan, (eds.), 2005, 53-66.

Cavalli-Sforza, Luigi & William S.-Y. Wang (1986). Spatial distance and lexical replacement. *Language* 62:38-55.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson (2005). Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science* 309-5743:2072-2075.

Felsenstein, Joseph (2003). *Inferring Phylogenies*. Sunderland, Mass.: Sinauer.

Gooskens, Charlotte (2004). Norwegian dialect distances geographically explained. In Britt-Louise Gunnarson, L.Bergström, G.Eklund, S.Fridella, L.H.Hansen, A.Karstadt, B.Nordberg, E.Sundgren, & M.Thelander, (eds.) *Language Variation in Europe. Papers from the 2nd International Conference on Language Variation in Europe (ICLaVE 2), June 12-14, 2003*, 195-206. Uppsala: Uppsala University.

Heeringa, Wilbert (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, University of Groningen.

Heeringa, Wilbert & Charlotte Gooskens (2003). Norwegian dialect examined perceptually and acoustically. *Computers and the Humanities* 37-3:293-315.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens, & John Nerbonne (2006). Evaluation of string distance algorithms for dialectology. In Nerbonne & Hinrichs (eds.), 2006, 51-62.

Heeringa, Wilbert & John Nerbonne (2001). Dialect areas and dialect continua. *Language Variation and Change* 13-3:375-400.

Heeringa, Wilbert & John Nerbonne (2006). De analyse van taalvariatie in het Nederlandse dialectgebied: methoden en resultaten op basis van lexicon en uitspraak. *Nederlandse Taalkunde* 11-3: 218-257.

Heeringa, Wilbert, John Nerbonne, & Peter Kleiweg (2002). Validating dialect comparison methods. W.Gaul & G.Ritter (eds.) *Proceedings of the 24th Annual Meeting of the Gesellschaft für Klassifikation,* 445-452. Heidelberg: Springer.

Hoppenbrouwers, Cor & Geer Hoppenbrouwers (1988). De featurefrequentiemethode en de classificatie van nederlandse dialecten. *TABU: Bulletin voor Taalwetenschap* 18-2:51-92.

Kessler, Brett (1995). Computational dialectology in Irish Gaelic. *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, 60-67. Dublin: ACL.

Kondrak, Grzegorz (2002). *Algorithms for Language Reconstruction*. PhD thesis, University of Toronto.

Kondrak, Grzegorz & Tarek Sherif (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. In Nerbonne & Hinrichs (eds.) 2006, 37-44.

Mace, Ruth, Clare J. Holden, & Stephen Shennan (eds.) (2005). *The Evolution of Cultural Diversity: A Phylogenetic Approach.* London: UCL Press.

Nakleh, Luay, Don Ringe, & Tandy Warnow (2005). Perfect phylogentic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81-2:382-420.

Mackay, Wesley & Grzegorz Kondrak (2005) . Comparing word similarity and identifying cognates with pair hidden Markov models. *Proceedings of the 9th Conference on Natural Language Learning (CoNLL),* 40-47. Shroudsburg, Penn.: ACL.

Nerbonne, John & Erhard Hinrichs (eds.) *Linguistic Distances.* Proc. of a workshop held at the joint meeting of ACL and COLING, Sydney, July, 2006. Shroudsburg, Penn.: ACL.

Nerbonne, John & Peter Kleiweg (forthcoming). Toward a dialectological yardstick. *Journal of Quantitative Linguistics.* 14.

Nerbonne, John & Wilbert Heeringa (1997). Measuring dialect distance phonetically. In John Coleman, (ed.) *Computational Phonology*. 11-18. Madrid: ACL.

Nerbonne, John, Wilbert Heeringa, Erik van den Hout, Peter van der Kooi, Simone Otten, & Willem van de Vis (1996). Phonetic distance between Dutch dialects. In Gert Durieux, Walter Daelemans, & Stephen Gillis (eds.), *CLIN VI: Proc. from the Sixth Computational Linguistics in the Netherlands Meeting.* 185-202. Antwerp: Centre for Dutch Language and Speech, University of Antwerp (UIA). Avail. at http://www.let.rug.nl/nerbonne/paper.html.

Nerbonne, John, Wilbert Heeringa, & Peter Kleiweg (1999). Edit distance and dialect proximity. In David Sankoff & Joseph Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* 2nd ed. v--xv. Stanford, Cal: CSLI

Vieregge, Wilhelm H., Toni C.M. Rietveld & Carel Jansen (1984). A distinctive feature based system for the evaluation of segmental transcription in Dutch. In Marcel P. R. van den Broecke & Antonie Cohen (eds.) *Proc. of the 10th International Congress of Phonetic Sciences*, 654-659. Dordrecht: Foris.