Computing and Historical Phonology

John Nerbonne
Alfa-Informatica
University of Groningen
j.nerbonne@rug.nl

T. Mark Ellison
Informatics
University of Western Australia
mark@markellison.net

Grzegorz Kondrak
Computing Science
University of Alberta
kondrak@cs.ualberta.ca

Abstract

We introduce the proceedings from the workshop 'Computing and Historical Phonology: 9th Meeting of the ACL Special Interest Group for Computational Morphology and Phonology'.

1 Background

Historical phonology is the study of how the sounds and sound systems of a language evolve, and includes research issues concerning the triggering of sound changes; their temporal and geographic propagation (including lexical diffusion); the regularity/irregularity of sound change, and its interaction with morphological change; the role of borrowing and analogy in sound change; the interaction of sound change with the phonemic system (potentially promoting certain changes, but also neutralizing phonemic distinctions); and the detection of these phenomena in historical documents.

There is a substantial and growing body of work applying computational techniques of various sorts to problems in historical phonology. We mention a few here to give the flavor of the sort of work we hoped to attract for presentation in a coherent SIGMORPHON workshop. Kessler (2001) estimates the likelihood of chance phonemic correspondences using permutation statistics; Kondrak (2002) develops algorithms to detect cognates and sound correspondences; McMahon and McMahon (2005) and also Nakhleh, Ringe and Warnow (2005) apply phylogenetic techniques to comparative reconstruction; and Ellison and Kirby (2006) suggest means of detecting relationships which do not depend on word by word comparisons. But we likewise wished to draw on the creativity of the computational linguistics (CL) community to see which other important problems in historical phonology might also be addressed computationally (see below).

There has recently been a good deal of computational work in historical linguistics involving phylogenetic inference, i.e., the inference to the genealogical tree which best explains the historical developments (Gray and Atkinson, 2003; Dunn et al., 2005). While the application of phylogenetic analysis has not universally been welcomed with open philological arms (Holm, 2007), it has attracted a good deal of attention, some of which we hoped to engage. We take no stand on these controversies here, but note that computing may be employed in historical linguistics, and in particular in historical phonology in a more versatile way, its uses extending well beyond phylogenetic inference.

2 Introduction

The workshop thus brings together researchers interested in applying computational techniques to problems in historical phonology. We deliberately defined the scope of the workshop broadly to include problems such as identifying spelling variants in older manuscripts, searching for cognates, hypothesizing and confirming sound changes and/or sound correspondences, modeling likely sound changes, the relation of synchronic social and geographic variation to historical change, the detection of phonetic signals of relatedness among potentially related languages, phylogenetic reconstruction based on sound correspondences among languages, dating

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 1–5,
Prague, June 2007.

historical changes, or others.

We were emphatically open to proposals to apply techniques from other areas to problems in historical phonology such as applying work on confusable product names to the modeling of likely sound correspondences or the application of phylogenetic analysis from evolutionary biology to the problem of phonological reconstruction.

3 Papers

We provide a preview to some of the issues in the papers in this bundle.

Brett Kessler's invited contribution sketches the opportunities for multiple string alignment, which would be extremely useful in historical phonology, but which is also technically so challenging that Gusfield (1999, Ch. 14) refers to it as "the holy grail" (of algorithms on strings, trees, and sequences).

3.1 Identification of Cognates

T. Mark Ellison combines Bayes's theorem with gradient descent in a method for finding cognates and correspondences. A formal model of language is extended to include the notion of parent languages, and a mechanism whereby parent languages project onto their descendents. This model allows the quantification of the probability of word lists in two languages given a common ancestor which was the source for some of the words. Bayes's theorem reverses this expression into the evaluation of possible parent languages. Gradient descent finds the best, or at least a good one, of these. The method is shown to find cognates in data from Russian and Polish.

Grzegorz Kondrak, David Beck and Philip Dilts apply algorithms for the identification of cognates and recurrent sound correspondences proposed by Kondrak (2002) to the Totonac-Tepehua family of indigenous languages in Mexico. Their long-term objective is providing tools for rapid construction of comparative dictionaries for relatively unfamiliar language families. They show that by combining expert linguistic knowledge with computational analysis, it is possible to quickly identify a large number of cognate sets across related languages. The experiments led to the creation of the initial version of an etymological dictionary. The authors hope that the dictionary will facilitate the reconstruction of a more accurate Totonac-Tepehua family tree, and shed light on the problem of the family origins and migratory patterns.

Michael Cysouw and Hagen Jung use an iterative process of alignment between words in different languages in an attempt to identify cognates. Instead of using consistently coded phonemic (or phonetic) transciption, they use practical orthographies, which has the advantage of being applicable without expensive and error-prone manual processing. Proceeding from semantically equivalent words in the Intercontinental Dictionary Series (IDS) database, the program aligns letters using a variant of edit distance that includes correspondences of one letter with two or more, ("multi-n-gram"). Once initial alignments are obtained, segment replacement costs are inferred. This process of alignment and inferring segment replacement costs may then be iterated. They succeed in distinguishing noise on the one hand from borrowings and cognates on the other, and the authors speculate about being able to distinguish inherited cognates from borrowings.

3.2 A View from Dialectology

Several papers examined language change from the point of view of dialectology. While the latter studies variation in space, the former studies variation over time.

Hans Goebl, the author of hundreds of papers applying quantitative analysis to the analysis of linguistic varieties in dialects, applies his dialectometric techniques both to modern material (1900) from the *Atlas Linguistique de France* and to material dating from approximate 1300 provided by Dutch Romanists. Dialectometry aims primarily at establishing the aggregate distances (or conversely, similarities), and Goebl's analysis shows that these have remain relatively constant even while the French language has changed a good deal. The suggestion is that geography is extremely influential.

Wilbert Heeringa and Brian Joseph first reconstruct a protolanguage based on Dutch dialect data, which they compare to the proto-Germanic found in a recent dictionary, demonstrating that their reconstruction is quite similar to the proto-Germanic, even though it is only based on a single branch of a large family. They then apply a variant of edit distance to the pronunciation of the protolanguage, comparing it to the pronunciation in modern Dutch dialects, allowing on the one hand a quantitative evaluation of the degree to which "proto-Dutch" correlates with proto-Germanic ($r = 0.87$), and a sketch of conservative vs. innovative dialect areas in the Netherlands on the other.

Anil Singh and Harshit Surana ask whether corpus-based measures can be used to compare languages. Most research has proceeded from the assumption that lists of word pairs be available, as indeed they normally are in the case of dialect atlas data or as they often may be obtained by constructing lexicalizations of the concepts in the so-called "Swadesh" list. But such data is not always available, nor is it straightforward to construct. Singh and Surana construct n-gram models of order five (5), and compare Indo-Iranian and Dravidian languages based on symmetric cross-entropy.

Martijn Wieling, Therese Leinonen and John Nerbonne apply PAIR HIDDEN MARKOV MODELS (PHMM), introduced to CL by Mackay and Kondrak (2005), to a large collection of Dutch dialect pronunciations in an effort to learn the degree of segment differentiation. Essentially the PHMM regards frequently aligned segments as more similar, and Wieling et al. show that the induced similarity indeed corresponds to phonetic similarity in the case of vowels, whose acoustic properties facilitate the assessment of similarity.

3.3 Views from other Perspectives

Several papers examined diachronic change from well-developed perspectives outside of historical linguistics, including evolution and genetic algorithms, language learning, biological cladistics, and the structure of vowel systems.

Monojit Choudhury, Vaibhav Jalan, Sudeshna Sarkar and Anupam Basu distinguish two components in language developments, on the one hand functional forces or constraints including ease of articulation, perceptual contrast, and learnability, which are modeled by the fitness function of a genetic algorithm (GA). On the other hand, these functional forces operate against the background of linguistic structure, which the authors dub 'genotype–phenotype mapping', and which is realized by the set of forms in a given paradigm, and a small set of possible atomic changes which map from form set to form set. They apply these ideas to morphological changes in dialects of Bengali, an agglutinative Indic language, and they are able to show that some modern dialects are optimal solutions to the functional constraints in the sense that any further changes would be worse with respect to at least one of the constraints.

Eric Smith applies the gradual learning algorithm (GLA) developed in Optimality Theory by Paul Boersma to the problem of reconstructing a dead language. In particular the GLA is deployed to deduce the phonological representations of a dead language, Elamite, from the orthography, where the orthography is treated as the surface representation and the phonological representation as the underlying representation. Elamite was spoken in southwestern and central Iran, and survives in texts dating from 2400– 360 BCE, written in a cuneiform script borrowed from Sumerians and Akkadians. Special attention is paid to the difficult mapping between orthography and phonology, and to OT's Lexicon Optimization module.

Antonella Gaillard-Corvaglia, Jean-Léo Léonard and Pierre Darlu apply cladistic analysis to dialect networks and language phyla, using the detailed information in phonetic changes to increase the resolution beyond what is possible with simple word lists. They examine Gallo-Romance vowels, southern Italo-Romance dialects and Mayan languages, foregoing analyses of relatedness based on global resemblance between languages, and aiming instead to view recurrent phonological changes as first-class entities in the analysis of historical phonology with the ambition of including the probability of specific linguistic changes in analyses.

Animesh Mukherjee, Monojit Choudhury, Anupam Basu and Niloy Ganguly examine the structure of vowel systems by defining a weighted network where vowels are represented by the nodes and the likelihood of vowels' co-occurring in the languages of the world by weighted edges between nodes. Using data from the 451 languages in the UCLA Phonological Segment Inventory Database (UPSID), Mukherjee and colleagues seek high-frequency symmetric triplets (with similar co-occurrence weights). The vowel networks which emerged tend to organize themselves to maximize contrast between the vowels when inventories are small, but they tend to grow by systematically applying the same contrasts (short vs long, oral vs nasal) across the board when they grow larger.

3.4 Methodology

Finally, there were three papers focusing on more general methodological issues, one on non-linearity, one on a direct manipulation interface to cross-tabulation, and one on visualizing distance measures.

Hermann Moisl has worked a great deal with the Newcastle Electronic Corpus of Tyneside English (NECTE). NECTE is a corpus of dialect speech from Tyneside in North-East England which was collected in an effort to represent not only geographical, but also social variation in speech. In the contribution to this volume, Moisl addresses the problem of nonlinearity in data, using the distribution of variance in the frequency of phonemes in NECTE as an example. He suggests techniques for spotting nonlinearity as well as techniques for analyzing data which contains it.

Tyler Peterson and Gessiane Picanco experiment with cross tabulation as an aid to phonemic reconstruction. In particular they use PIVOT TABLES, which are cross tabulations supported by new database packages, and which allow direct manipulation, e.g., drag and drop methods of adding and removing new sets of data, including columns or rows. This makes it easier for the linguist to track e.g. phoneme correspondences and develop hypotheses about them. Tupí stock is a South American language family with about 60 members, mostly in Brazil, but also in Bolivia and Paraguay. Pivot tables were employed to examine this data, which resulted in a reconstruction a great deal like the only published reconstruction, but which nevertheless suggested new possibilities.

Thomas Pilz, Axel Philipsenburg and Wolfram Luther describe the development and use of an interface for visually evaluating distance measures. Using the problem of identifying intended modern spellings from manuscript spellings using various techniques, including edit distance, they note examples where the same distance measure performs well on one set of manuscripts but poorly on another. This motivates the need for easy evaluation of such measures. The authors use multidimensional scaling plots, histograms and tables to expose different levels of overview and detail.

3.5 Other

Although this meeting of SIGMORPHON focused on contributions to historical phonology, there was also one paper on synchronic morphology.

Christian Monson, Alon Lavie, Jaime Carbonell and Lori Levin describe ParaMor, a system aimed at minimally supervised morphological analysis that uses inflectional paradigms as its key concept. ParaMor gathers sets of suffixes and stems that co-occur, collecting each set of suffixes into a potential inflectional paradigm. These candidate paradigms then need to be compared and filtered to obtain a minimal set of paradigms. Since there are many hundreds of languages for which paradigm discovery would be a very useful tool, ParaMor may be interesting to researchers involved in language documentation. This paper sketches the authors' approach to the problem and presents evidence for good performance in Spanish and German.

4 Prospects

As pleasing as it to hear of the progress reported on in this volume, it is clear that there is a great deal of interesting work ahead for those interested in computing and historical phonology. This is immediately clear if one compares the list of potential topics noted in Sections 1-2 with the paper topics actually covered, e.g. by skimming Section 3 or the table of contents. For example we did not receive submissions on the treatment of older documents, on recognizing spelling variants, or on dating historical changes.

In addition interesting topics may just now be rising above the implementation horizon, e.g. computational techniques which strive to mimic internal reconstruction (Hock and Joseph, 1996), or those which aim at characterizing general sound changes, or perspectives which attempt to tease apart historical, areal and typological effects (Nerbonne, 2007). In short, we are optimistic about interest in follow-up workshops!

5 Acknowledgments

We are indebted to our program committee, to the incidental reviewers named in the organizational section of the book, and to some reviewers who remain anonymous. We also thank the SIGMORPHON chair Jason Eisner and secretary Richard Wicentowski for facilitating our organization of this workshop under the aegis of SIGMORPHON, the special interest group in morphology and phonology of the Association for Computational Linguistics.[1] We thank Peter Kleiweg for managing the production of the book. We are indebted to the Netherlands Organization for Scientific Research (NWO), grant 235-89-001, for cooperation between the Center for Language and Cognition, Groningen, and the *Department of Linguistics* The Ohio State University, for support of the work which is reported on here.

References

A. Michael Dunn, A. Terrill, Geert Reesink, and Stephen Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

Mark Ellison and Simon Kirby. 2006. Measuring language divergence by intra-lexical divergence. In *Proc. of ACL/COLING 2006*, pages 273–280, Shroudsburg, PA. ACL.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Dan Gusfield. 1999. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge Univesity Press, Cambridge.

Hans Henrich Hock and Brian D. Joseph. 1996. *Language History, Language Change, and Language Relationship : An Introduction to Historical and Comparative Linguistics*. Mouton de Gruyter, Berlin.

Hans J. Holm. 2007. The new arboretum of Indo-European “trees”: Can new algorithms reveal the phylogeny and even prehistory of IE? *Journal of Quantitative Linguistics*, 14(2).

Brett Kessler. 2001. *The Significance of Word Lists*. CSLI Press, Stanford.

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

Wesley Mackay and Grzegorz Kondrak. 2005. Comparing word similarity and identifying cognates with pair hidden markov models. In *Proceedings of the 9th Conference on Natural Language Learning (CoNLL)*, pages 40–47, Shroudsburg, PA. ACL.

April McMahon and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford University Press, Oxford.

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

John Nerbonne. 2007. Review of April McMahon & Robert McMahon *Language Classification by Numbers*. Oxford: OUP, 2005. *Linguistic Typology*, 11.

[1] `http://nlp.cs.swarthmore.edu/sigphon/`