Review of *Statistics in Historical Linguistics*
Stephen Grimes
November 19, 2001

Embleton, Sheila M. (1986). Statistics in Historical Linguistics. Bochum, Studienverlag Brockmeyer.

Introduction

In this 1986 book, Sheila Embleton surveys the development and use of statistical techniques in historical linguistics. She argues that any successful lexicostatistic model must incorporate word borrowing rates. A computational model for family tree reconstruction using borrowing is introduced, building on ideas of Sankoff (Sankoff 1972). Embleton outlines future uses of lexicostatistics and suggests much work is yet to be done.

Synopsis

I begin with a clarification of the terms lexicostatistics and glottochronology. Embleton takes lexicostatistics to refer to statistical techniques that measure the degree of closeness of two genetically related languages. Glottochronology is a specific method for calculating the time depth of a posited ancestral language.

The first two chapters are devoted to giving a background of lexicostatistical methods. Chapter one discusses techniques developed before the 1960's to measure the degree of similarity between languages. Many of the models discussed here seem to have critical flaws and most are no longer discussed actively in the literature. By addressing these older methods early on, Embleton shows how far lexicostatistics has been developed since its early roots. The first chapter serves as a good reference for the lexicostatistics of the first half of the 20th century.

Chapter two is devoted entirely to Swadesh's glottochronology approach in the 1950's and the discussion that ensued. Embleton classifies papers on glottochronology into three categories: refuting, supporting, and neutral. Strangely, neutral papers typically assume glottochronology is a valid theory and apply it to get results for a specific language family. Good examples of the debate surrounding glottochronology can be also be found in Hymes (Hymes 1960).

While the first two chapters are expository in nature, the remaining six chapters are devoted to developing a lexicostatistical model that Embleton sees as being most promising for the future. The most central theme developed is that a successful lexicostatistical approach must incorporate borrowing rates, and Embleton provides several illustrations to show glottochronology results are often skewed precisely because they have failed to account for borrowing. Much of Embleton's work here originally appeared as part of her unpublished Ph.D. thesis at University of Toronto (Embleton 1981).

The third and fourth chapters provide detailed derivations of the mathematics behind her improvement to Sankoff's borrowing model. She corrects important but seemingly obvious oversights in some of Sankoff's formulas, then tests her improved model on computer-simulated data. Testing her model against hypothetical data is

certainly an important prerequisite, as it shows the model would make an accurate prediction of a language that conforms to her simplifying assumptions.

Chapters five through seven discuss applications of the statistical borrowing model to Germanic, Romance, and Wakashan. These results were also previously published(Embleton 1985). An abundance of historical and linguistic data exists for Germanic and Romance, so these language families serve as test cases for the computational model Embleton has developed.

Embleton's model is very accurate in predicting the correct topology of the family tree in the Germanic and Romance cases, though skeptics might argue that it was developed specifically for Indo-European and may fail in other language families. Also, Embleton finds time-depths predicted by her model reflect accurately the independently predicted time depths using historical, archeological, or other linguistic information. Embleton notes that when the model predicts an inaccurate separation date, the actual time split is usually earlier than the model predicts. Embleton has a convenient explanation in each case, usually that early, undetected borrowings between the two languages pairs went unnoticed a corrected borrowing rate would solve the problem. Embleton highlights the case of the conservative Icelandic, a language that previously did not appear to conform well to a glottochronological analysis (Bergsland and Vogt, 1962).

Scholars do not agree about the exact tree structure for Wakashan and there is little evidence to verify time depth predictions made by Embleton's model in the Wakashan case. Embleton notes that this very situation is where lexicostatistics excels: it makes a statistical prediction about a *tentative* family tree, thereby giving historical linguists a working hypothesis where none had previously existed. Embleton chose Wakashan because scholars note heavy borrowing due to close geographic proximity and "potlatch" relationships, and notes that the borrowing model makes significantly better predictions than the traditional glottochronological model.

Embleton concludes the book by making some general statements about the state of lexicostatistics. Embleton herself has been criticized for being both too harsh and too supportive in her views of glottochronology. Embleton calls for more well documented case studies to be conducted, and that future work be supervised closely by trained mathematicians. She implies that lexicostatistics has received undue harsh criticism, and that models of phonology and syntax have not been adequately justified either. She cautions that results of glottochronology are valuable in some contexts, but must be interpreted correctly by scholars already familiar with the given language family.

Critical Analysis

Embleton succeeds in the difficult and awkward task of surveying past lexicostatistic methods, which range from being insightful to fatally flawed. She also does an excellent job narrating the glottochronology debate, which requires her to walk the fine line of being supportive and encouraging of further research while appearing rational and critical of some of the generalizations made in this field.

There are two points I especially thought were argued well in the book. First, as mentioned several times above, she clearly motivates that any *potentially* successful lexicostatistical model must incorporate borrowing. Many scholars refute classical glottochronology because it does not address the causes of language decay, which are numerous. By recognizing that language contact affects language change, Embleton is

addressing one of lexicostatistics' most prominent concerns and moving toward a more unified approach.

Embleton also softens the claim that glottochronology makes by pointing out that it was never intended to be a deterministic model. With each projected time depth must come a confidence interval, a range of dates over which the expected separation date must lie. I feel it is an important distinction, but one begins to wonder what future Embleton sees in lexicostatistics when she places so many disclaimers on the method's application.

There are two major problems with this book. The first is that it reads more like a lengthy journal article rather than serving as a reference on historical linguistics and statistics. The majority of the book aims to present and evaluate the current computational model being proposed by Embleton, and the first two chapters are simply a lengthy preface to this discussion. We do not see an alternative approach discussed. Unfortunately, the fault of this may not rest entirely on Embleton, as lexicostatistics may be suffering from a lack of recent innovation.

Embleton has difficulty presenting an unbiased view of statistics in historical linguistics. She gives voice to both supporters and detractors alike, attempting to provide answers to concerns raised against lexicostatistics while at the same time admitting its faults. However, because Embleton believes in the possibility of developing a successful model, she fails to acknowledge or adequately address all of her underlying assumptions. For instance, one of the severe problems of glottochronology has always been the assumption that languages change at more or less a constant rate (Swadesh 1950; Swadesh 1952). On one hand Embleton seems to concede that *a priori* there is not reason to think two given languages should change at a constant pace. Surprisingly, however, her new model does presume that, for a given language, the replacement rate (and indeed the borrowing rate as well) will be constant over a period of time!

A more minor point that Embleton quickly glosses over is the status family tree model. It is natural to employ family trees to describe genetic classification, but there are other approaches (Campbell 1999). Her parameterized model itself suggests in the future more of a wave theory approach to language evolution, where each word has a given probability of changing or being replaced over time. By not giving more attention to assumptions of the family tree model, one might imagine Embleton missed a chance to interact with recent work on dialectology and work towards a more unified approach.

Concluding Remarks

This book is relevant to all historical linguists and statisticians interested in historical linguistics. Linguists without formal mathematics and statistics training may find some sections daunting, but fortunately most of the central debates in lexicostatistics surround its actual assumptions, not the mathematics.

The book is a necessary resource for the linguist or anthropologist interested in developing theories, applying models, or interpreting the findings of lexicostatistics and glottochronology. The work functions to level the playing field for future research and avoid duplication of developments and criticisms. It injects several new ideas into the field while being conservative about the usefulness of the method.

References

Campbell, L. (1999). Historical Linguistics, an Introduction. Cambridge, MIT Press.

Embleton, S. M. (1981). Incorporating borrowing rates in lexicostatistical tree reconstruction, Univeristy of Toronto.

Embleton, S. M. (1985). "Lexicostatistics applied to the Germanic, Romance, and Wakashan families." Word **36**: 37-60.

Hymes, D. H. (1960). "Lexicostatistics so far." Current Anthropology **1**(1): 3-44.

Sankoff, D. (1972). "Reconstructing the history and geography of an evolutionary tree." American Mathematics Monthly **79**: 596-603.

Swadesh, M. (1950). "Salish internal relationships." Internation Journal of American Linguistics **16**: 157-167.

Swadesh, M. (1952). "Lexico-statistic dating of prehistoric ethnic contacts." Proceedings of the American Philosophical Society **96**: 452-463.