Abstract for The Fourteenth International Symposium on Malay/Indonesian Linguistics (ISMIL 14), Minneapolis, Minnesota, USA, 30 April – 2 May 2010

Title
Malay varieties of southern Sumatra: An evaluation of Levenshtein distance as a tool for dialect classification

Authors
Eldwin Truong
Language assessment specialist
Indonesia Survey Team
SIL International Indonesia

Contact Person
Eldwin Truong
SIL International Indonesia
PO Box 1561/JKS
Jakarta 12015
Indonesia
eldwin_truong@sil.org
Office: +62.21.7581.6425
Mobile: +62.813.1134.7934

Abstract
Phonological analysis of dialect variation has traditionally relied on qualitative methods that are time-intensive and require many individual subjective judgments on the part of the researcher. In historical-comparative analysis, for example, relationships between linguistic varieties are reconstructed based on consistent sound changes (Campbell, 1998). An additional quantitative computational method of analysis would aid in more quickly determining dialect groupings for areas where language varieties have not been thoroughly studied. This paper compares a computational analysis using Levenshtein distance (Kessler, 1995; Nerbonne and Heeringa, 1996) to an application of the historical-comparative method for determining dialect variation. The analyses are carried out on the previously understudied Malay varieties of South Sumatra and Bengkulu provinces on the island of Sumatra, Indonesia.

Levenshtein distance is a measure of of the distance between two strings, which can be used as a quantitative method for phonetic analysis of dialect variation. This application of Levenshtein distance measures the phonetic distance between lexical items in wordlists in different dialects by computing the minimum number of changes required to convert one lexical string to another. The average distance for all pairs of lexical items between wordlists is calculated, and the resulting measures are compared to show the relative distance between varieties (Nerbonne and Heeringa, 1996). This differs from lexicostatistics, which counts only lexical similarity and ignores phonetic difference. Analyses utilizing Levenshtein distance have recently been applied to study already well-

documented dialect situations in Europe, but it have not yet been widely applied to understudied languages outside of Europe.

The data analyzed in this paper consists of approximately 90 wordlists of 358 items which were collected throughout South Sumatra and Bengkulu provinces over a two-year period. McDowell and Anderbeck (forthcoming) conducted a historical-comparative analysis of the data to track sound changes from Proto-Malayic and posit dialect groupings based on these sound changes. Here, the researcher analyzes the same data using Levenshtein distance.

The results from both the historical-comparative analysis and Levenshtein distance analysis show two main groups of Malay varieties in southern Sumatra. The Levenstein distance analysis produces grouping results very similar to those of the historical-comparative analysis, supporting the classification of these two groups as separate languages. Both analyses also show several subgroups of language varieties under each of the main groupings but the subgroupings are not identical. The differences in subgrouping classification between the two analyses and their significance are discussed, along with possible explanations.

This comparison shows that computational analysis using Levenshtein distance is able to provide useful results for dialectology research. The addition of Levenshtein distance analysis to more established research methods will help to improve dialect classification in previously understudied languages.

References

Campbell, Lyle. 1998. *Historical Linguistics*. Cambridge, Massachusetts: MIT Press.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics (ACL)*. Dublin. 60–67.

McDowell, Jonathan & Karl Anderbeck. Forthcoming. Bhinnêka tunggal ika: Unity in diversity. Malayic varieties of southern Sumatra.

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten & Willem van de Vis. 1996. Phonetic Distance between Dutch Dialects. In G. Durieux, W. Daelemans and S. Gillis (eds.), *CLIN VI: Proceedings of the Sixth CLIN Meeting*. Antwerp: Centre for Dutch Language and Speech (UIA). 185-202.