Data linguistica 27

Studies in computational historical linguistics

av Taraka Rama

Akademisk avhandling för filosofie doktorsexamen i språkvetenskaplig databehandling,
som enligt beslut av humanistiska fakultetsnämnden vid
Göteborgs universitet kommer att försvaras offentligt fredagen den
13 november 2015 kl. 13.15 i Lilla hörsalen, Humanisten.

GÖTEBORGS UNIVERSITET
HUMANISTISKA FAKULTETEN

Göteborg 2015

TITLE: Studies in computational historical linguistics
LANGUAGE: English
AUTHOR: Taraka Rama

Abstract

Computational analysis of historical and typological data has made great progress in the last fifteen years. In this thesis, I work with vocabulary lists for addressing some classical problems in historical linguistics such as cognate identification, discriminating related languages from unrelated languages, assigning possible dates to splits in a language family, and providing an internal structure to a language family. I compare the internal structure inferred from vocabulary lists with the family trees given in Ethnologue. I explore the ranking of lexical items in the widely used Swadesh word list and compare my ranking to another quantitative reranking method and short word lists composed for discovering long-distance genetic relationships. I show that the choice of string similarity measures is important for internal classification and for discriminating related from unrelated languages. The dating system presented in this thesis can be used for assigning age estimates to any new language group and overcomes the assumption of a constant rate of lexical replacement assumed by glottochronology. I train and test a linear classifier based on gap-weighted subsequence features for the purpose of cognate identification. An important conclusion from these results is that n-gram approaches can be used for different historical linguistic purposes.

KEYWORDS: Automatic language classification, calibration dates, cognate identification, computational historical linguistics, internal classification, language families, n-grams, skip-grams, string similarity measures, typological data, word lists.

DISTRIBUTION:
Department of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

Data linguistica 27
ISSN 0347-948X
ISBN 978-91-87850-58-5
E-publication http://hdl.handle.net/2077/40571

PRINTED in Sweden by Taberg Media Group AB Göteborg 2015