CHALMERS |  UNIVERSITY OF GOTHENBURG

Unsupervised Learning of Morphology and the Languages of the World

Harald Hammarström

Thesis for the Degree of Doctor of Engineering
to be defended in public at **10:15, December 11, 2009**
in room HB1, Hörsalsvägen 8,
Chalmers University of Technology (Campus Johanneberg), Gothenburg.

The defence will be held in English.
Faculty opponent: Professor Richard Sproat,
Oregon Health and Science University.

Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg, Sweden
Telephone: +46 (0)31 772 1000

Unsupervised Learning of Morphology and the Languages of the World

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
ISBN 978-91-628-7942-6

Abstract

This thesis presents work in two areas; Language Technology and Linguistic Typology.

In the field of Language Technology, a specific problem is addressed: *Can a computer extract a description of word conjugation in a natural language using only written text in the language?* The problem is often referred to as Unsupervised Learning of Morphology and has a variety of applications, including Machine Translation, Document Categorization and Information Retrieval. The problem is also relevant for linguistic theory. We give a comprehensive survey of work done so far on the problem and then describe a new approach to the problem as well as a number of applications. The idea is that concatenative affixation, i.e., how stems and affixes are stringed together to form words, can, with some success, be modelled simplistically. Essentially, words consist of high-frequency strings ("affixes") attached to low-frequency strings ("stems"), e.g., as in the English *play-ing*. Case studies show how this naive model can be used for stemming, language identification and bootstrapping language description.

There are around 7 000 languages in the world, exhibiting a bewildering structural diversity. Linguistic Typology is the subfield of linguistics that ams to understand this diversity. Many of the languages in the world today are spoken only by relatively small groups of people and are threatened by extinction and it is therefore a priority to record them. Language documentation, is and has been, an extremely decentralised activity, carried out not only by linguists, but also missionaries, travellers, anthropologists etc foremostly throughout the past 200 years. There is no central record of which and how many languages have been described. To meet the priority, we have attempted to list those languages which are the most poorly described which do not belong to a language family where some other languages is decently described – a task requiring both analysis and diligence. Next, the thesis includes typological work on one of the more tractable aspects of language structure, namely numeral systems, i.e., normed expressions used to denote exact quantities. In one of the first surveys to cover the whole world, we look at rare number bases among numeral systems. One major rarity is base-6-36 systems which are only attested in South/Southwest New Guinea and we make a special inquiry into its emergence.

Traditionally, linguists have had headaches over what counts as a language as opposed to a dialect, and have therefore been reluctant to give counts of the number of languages in a given area. One chapter of the present thesis shows that, contrary to popular belief, there is an intuitively sound way to count languages (as opposed to dialects). The only requirement is that, for each pair of varieties, we are told whether they are mutually intelligible or not.