

Abstract of thesis entitled

Multilingual Information Retrieval on the World Wide Web:

The Development of a Cantonese-Dagaare-English Trilingual Electronic Lexicon

Submitted by

Mok Yuen Kwan Sally

for the degree of Master of Philosophy
at the University of Hong Kong
in May 2006

The World Wide Web (WWW) contains some constraints on the manipulation of multilingual textual data. In the literature on computational lexicography, studies have analysed the design and compilation of electronic dictionaries, but not much research has been done on language resources that handle multilingual lexical transfer and natural language processing (NLP) on the internet.

This thesis examines storage and retrieval problems in developing a Cantonese-Dagaare-English trilingual online lexicon which is human-readable (Schryver, 2003). It investigates the following computational and linguistic issues concerning web internationalization: the language input, scripting languages and database systems, font systems, character encodings and text normalization. It further suggests a factorial approach to control multilingual data retrieval on the WWW. It raises the question as to whether the WWW can be considered as a truly World Wide application for global



communications with reference to the PLT (People, Languages, Tools) conditions proposed.

In the end, the thesis discusses the challenges for building an NLP Cantonese-Dagaare-English lexicon by studying the existing multilingual database models used in two NLP lexical frameworks, WordNet (Beckwith & Miller et al., 1993) and Papillion (Mangeot-Lerebours & Serasset, 2002). This analysis will shed light on the future directions towards developing an NLP lexicon for Dagaare, Cantonese and other languages not included in the study.

Word-count: 209 words

