Building Language Resources: Ways to move forward

Anne David and Michael Maxwell
Center for Advanced Study of Language
University of Maryland, USA
aeadavid@gmail.com, maxwell@umiacs.umd.edu

Abstract

There are perhaps seven thousand languages in the world, ranging from the largest with hundreds of millions of speakers, to the smallest, with one speaker. On a different axis, languages can be ranked according to the quantity and quality of computational resources. Not surprisingly, there are correlations between these two axes: languages like English and Mandarin have substantial resources, while many of the smallest languages are completely undocumented. Nevertheless, the correlation is not perfect; there are languages with a million speakers which are more or less unwritten, and there are very large languages – some of the languages of India, for example – which are relatively resource-poor.

Unfortunately, what counts as resource-rich (or even resource-adequate) in computational linguistics is a moving target. For languages to move in the direction of resource richness, considerable effort (people and money) have to be provided over a prolonged period of time. One can sit back and wait for this to happen, or give up; alternatively, one can map out a realistic way forward, building on the strengths of each language's situation.

Among the strengths which may prove useful to building computational resources for languages are the following:

- Long traditions of grammatical and lexical description
- Traditions of literacy and literature
- Local expertise in linguistics and computing
- The world-wide community of linguists and computer experts
- Resource availability in related languages

At the same time, there are weaknesses and other problems – some language specific, some more general – which need to be considered:

- Lack of consensus on ways of representing the language (scripts, character encoding)
- Complexities inherent in particular languages (complex scripts, complex morphologies, variant orthographies, diglossia, dialectal variation)
- Economic and educational realities in the countries where the language is spoken
- Political attitudes towards some languages, particularly minority languages
- The 'not invented here' syndrome
- Software obsolescence, and the potential obsolescence of language data

This talk will look at ways in which the strengths enumerated above might be leveraged, while avoiding the potential weaknesses.

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 1–2,
Hyderabad, India, January 2008.