Using the Rosetta Wordlists and LoCodes Database to Locate Samples of Low-Density Languages on the World Wide Web

Andrew MacKinlay

May 31, 2005

Introduction

The utility of language samples to anyone with an interest in a given language is obvious – they can be valuable to linguists, language technologists and speakers of the language, among others. The World Wide Web (WWW) is vast repository of information with potentially large numbers of samples of many languages, but often locating these samples reliably is a non-trivial task. While language-specific search tools on search engines such as Google are useful for the languages they cover, for the vast majority of languages which have fewer speakers and a smaller online presence (low-density languages) they provide no information, and it these lesser-known languages for which language resources are likely to be most useful to language researchers.

There are existing online resources which provide varying degrees of coverage for a large number of languages but the amount of data stored in these for even the best covered languages is

normalisation could be carried out to account for biases in the data (The equivalent of ‘article’ appears in each document about 50 times, yet is unlikely to be anywhere near the most likely