Relish: rendering endangered languages lexicons interoperable through standards harmonisation

Marc Kemps-Snijders

Max Planck Institute for Psycholinguistics
PO Box 310
6500 AH Nijmegen
The Netherlands

When a lexicon constitutes the only record of a dying or already extinct language, it can contribute unique linguistic and cultural information to our store of scientific knowledge. And making it interoperable with other lexical data becomes a critical research priority. However, despite the support accorded to initiatives to develop digital standards for language documentation within both the US and Germany, there still exist major barriers to lexicon interoperability. The most significant barrier is that standards-setting bodies have arrived at different standards for format and markup on the two sides of the Atlantic. On the European side the main focus has been towards the ISO 24623 Lexical Markup Framework (LMF) and the ISO 12620 Data Category Registry (DCR) while at the American side the Lexicon Interchange Format (LIFT) and GOLD have been the centre of attention. As a consequence, within each national community, divergences exist in lexicon format and markup, in part because field linguists have hitherto relied on software which does not offer the linguist adequate support in choosing structural or linguistic categories.

The Relish project will promote language-oriented research by addressing a two-pronged problem: (1) the lack of harmonization between digital standards for lexical information in Europe and America, and (2) the lack of interoperability among existing lexicons of endangered languages, in particular those created with the Shoebox lexicon building software. Focusing on six to eight lexicons of endangered languages, the project will establish a unified way of referencing lexicon structure and linguistic concepts, and develop a procedure for migrating these heterogeneous lexicons to a standards-compliant format. Once developed, the procedure will be generalizable to the large store of lexical resources involved in the LEGO and DoBeS projects. The project will produce significant benefits both to the user community and to the organizations which support their research.

As a first step the linguistic concepts expressed in GOLD will be harmonized with those already present in the Data Category Registry thus providing a unified and persistent reference point for concepts used on both sides of the Atlantic. Also, the most commonly used Shoebox markers for the Multi Dictionary Formatter (MDF) will be made available as data categories in the Data Category Registry to provide further support for lexica created using the Shoebox tool. Focusing on six to eight lexicons of endangered languages, the project will establish a unified way of referencing lexicon structure and linguistic concepts, and develop a procedure for migrating these heterogeneous lexicons to a standards-compliant format. To complement this bottom-up approach the Relish project uses a top-down approach analyzing existing standards for lexical resources (GOLD/LIFT and DCR/LMF) to identify commonalities and differences at the conceptual and structural level. An attempt is made to harmonize these standard approaches to come to a single interchange format making it possible to exchange lexica in a unified manner. Existing software tools as LEXUS and SOLID will be modified to support the interchange scenarios.

Kepa Sarasola, Francis M. Tyers, Mikel L. Forcada (eds.)
7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, Valetta, Malta, 23 May 2010, p. 1