Lexical Markup Framework (LMF):

working to reach a consensual ISO standard on lexicons

Tutorial presented by:

Laurent Romary (INRIA-Loria+CNRS: ISO-TC37/SC4 chairman)

Gil Francopoulo (INRIA-Loria: ISO-LMF co-editor)

Monica Monachini (CNR-ILC: NLP lexicons specialist)

Susanne Salmon-Alt (CNRS-ATILF: NLP lexicons specialist)

Four years ago, during LREC-2002, the ISO-TC37 National delegations decided to address standards dedicated to NLP. This was new. Up until now, not any ISO standard for NLP did exist.

These standards are currently elaborated as high level specifications and deal with word segmentation, annotations, feature structures and lexicons. These high level standards are based on low level specifications dedicated to constants, namely data categories (revision of ISO 12620), language codes, code scripts, country codes and Unicode. This work is in progress. And the two level organization will form a coherent family of standards.

Concerning lexicons, the standard is called "Lexical Markup Framework" (ISO 24613). The task is not easy because a good consensus must be reached and the scope is rather broad. First of all, a great number of hours of discussions and research has been devoted to specify the terminology and find (we hope) good definitions for the terms: lexicon, word, lemma, morpheme, affix, multi-word expression, syntactic behavior, sense, transfer etc. Let's note that the difficulty is not to write a good definition for each of these terms but rather to write a good coherent set of definitions for all these terms.

Types of individual instanciations of LMF can include monolingual, bilingual or multilingual resources. The range of targeted NLP applications is not restricted. The same specifications are to be used for both small and large lexicons. The descriptions range from morphology, syntax, semantic to translation. Special attention has been taken to multi-word expressions, being flexible, semi-flexible or fixed expressions, relying or not on the grammar of the given language. And if needed, the lexicons must be correctly linked to semantic web ontologies. The covered languages are not restricted to European languages but cover all natural languages, including Semitic and Asian languages. Special attention has been taken for languages with complex morphology and with languages with multiple orthographies, like a certain number of Asian languages.