Interoperability for linguistic and terminological content description

Wim Peters
Department of Computer Science
University of Sheffield
w.peters@dcs.shef.ac.uk

Linguistic knowledge is expressed in various ways in terminological and linguistic resources. The nature and format of this knowledge is determined by a number of factors, such as user needs and the required level of adhesion to existing standards for the representation of the linguistic knowledge. From a practical point of view, linguistic and terminological standards are in daily use for the purpose of resource creation (term banks, dictionaries, translation memories etc.). These different application areas are often unaware of cross-border standards and best practices. Therefore, given the existence of this variety of (standard) linguistic models, it is necessary to establish interoperability between their vocabularies in a principled way in order to enable interdisciplinary re-use and comparison.

There are many proposed standards and best practices for encoding linguistic and terminological knowledge. There are terminological models, such as the ISO standard initiatives ISO16620 and linguistic specifications, like the ISO 24613 Lexical Markup Framework (LMF) (Francopoulo et al., 2006). The ISOCAT initiative (Kemps-Snijders et al., 2008) aims to gather and integrate these models and create new standards. In the ontology area, ontologies for modelling linguistic knowledge, such as the Linguistic Information Repository (LIR) (Peters et al., 2007), capture various parts of the linguistic descriptive domain. Their purpose is to associate multilingual linguistic knowledge with conceptual ontology elements. In the translation memory area standards such as TMX[1] (Translation Memory eXchange), XLIFF[2] (XML Localization Interchange File Format) and MLIF[3] (Multi Lingual Information Framework) are widely used.

The present-day situation is that (standard) representation formats are non-exhaustive: they do not cover all aspects of linguistic description to the highest possible level of granularity. For instance, LMF is quite underspecified in its definitions of class attributes. Most models are partially overlapping and/or complementary, and there is no mechanism yet to formally capture the commonalities and differences between descriptive systems. Interoperability can be obtained by defining mappings between these models that allows the detection of overlap, complementarity and navigation through and cross resources.

I will present LingNet[4], a model for mapping linguistic and terminological (standard) information, which covers a.o. the following standards and best practices from the ontological, linguistic, terminological and translation memory disciplines: LIR, LMF, XLIFF, TMX, MLIF. This distributed mapping avoids localized inclusion of whole ontologies, and favours a more modular approach. The flexible mappings LingNet provides allow the identification of conceptually coherent building blocks that cover the exact descriptive requirements of the user, enabling a pick-and-mix combination of elements from models according to criteria of coverage, complementarity and granularity of linguistic description.

References

Francopoulo, G., Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria. 2006, LMF for multilingual, specialized lexicons. In Proc. of the LREC 2006 in Genova, Italy.

Kemps-Snijders, M. Windhouwer, M.E. Wittenburg, P. Wright, S.E., 2008, ISOcat: A Revised ISO TC 37 Data Category Registry. Presentation at the Conference on Terminology and Information Interoperability - Management of Knowledge and Content (TII 2008), Moscow, Russia

Peters, W. Montiel-Ponsoda, E. Aguado de Cea, G., 2007, Localizing Ontologies in OWL. In: Proceedings of the ISWC07 OntoLex workshop, Busan, Korea

[1] http://www.lisa.org/tmx/
[2] http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.pdf
[3] http://mlif.loria.fr/
[4] first version: http://www.gate.ac.uk/ns/ontologies/LingNet/LingNet-v0.1.owl.