Standards and best practice for multilingual computational lexicons:

ISLE MILE …. and more

Nicoletta Calzolari[1], Ralph Grishman[2], Martha Palmer[3]

[1]Istituto di linguistica Computazionale del CNR, Pisa, Italy
glottolo@ilc.cnr.it
[2]New York University
grishman@cs.nyu.edu
[3]Penn University
mpalmer@cis.upenn.edu

Abstract

ISLE (*International Standards for Language Engineering*) is a transatlantic standards oriented initiative under the Human Language Technology (HLT) programme within the EU-US International Research Co-operation. It is a continuation of the European EAGLES (*Expert Advisory Group for Language Engineering Standards)* initiative, carried out through a number of subsequent projects funded by the European Commission (EC) since 1993. Within the *multilingual computational lexicons* Working Group, ISLE aims at: extending EAGLES work on lexical semantics, necessary to establish inter-language links; designing and proposing standards for multilingual lexicons; developing a prototype tool to implement lexicon guidelines and standards; creating exemplary EAGLES-conformant sample lexicons and tagging exemplary corpora for validation purposes; and developing standardised evaluation procedures for lexicons. After a short introduction on the ISLE proposal for standards, the MILE (Multilingual ISLE Lexical Entry), we will focus the discussion on short and medium term requirements with respect to standards for multilingual lexicons and content encoding, in particular industrial requirements. We will stress the importance of reaching consensus on (linguistic and non-linguistic) "content", in addition to agreement on formats and encoding issues, and will define further steps necessary to converge on common priorities. Semantic Web standards and the needs of content processing technologies will be also addressed.

1. Goals of the Panel

ISLE[1] *International Standards for Language Engineering*) is a transatlantic standards oriented initiative under the Human Language Technology (HLT) programme within the EU-US International Research Co-operation. It is a continuation of the long standing European EAGLES (*Expert Advisory Group for Language Engineering Standards)* initiative, carried out through a number of subsequent projects funded by the European Commission (EC) since 1993.

Within the *multilingual computational lexicons* Working Group (CLWG), ISLE aims at: extending EAGLES work on lexical semantics, necessary to establish inter-language links; designing and proposing standards for multilingual lexicons; developing a prototype tool to implement lexicon guidelines and standards; creating exemplary EAGLES-conformant sample lexicons and tagging exemplary corpora for validation purposes; and developing standardised evaluation procedures for lexicons. The CLWG is committed to the consensual definition of a standardized infrastructure to develop multilingual resources for HLT applications, with particular attention to the needs of Machine Translation and Crosslingual Information Retrieval systems.

The Panel will include, in addition to ISLE members, developers and users of multilingual systems and of content management systems, and researchers interested in multilingual and content encoding standards.

After a short introduction on the ISLE proposal for the MILE (*Multilingual ISLE Lexical Entry*) - a general schema for the encoding of multilingual lexical information to be intended as a meta-entry, acting as a common representational layer for multilingual lexical resources -, we will focus the discussion on short and medium term requirements with respect to standards for multilingual lexicons and content encoding, in particular industrial requirements. We will stress the importance of reaching consensus on (linguistic and non-linguistic) "content", in addition to agreement on formats and encoding issues, and will try to define further steps necessary to converge on common priorities. Semantic Web standards and the needs of content processing technologies will be also addressed.

2. A few Issues for the Panel

If we break the global problem of multilingual content technologies into small more manageable pieces, Linguistic Resources (LR) are certainly one of these pieces. Which is the relevance and impact of the availability of (good, deep, knowledge intensive) resources (lexicons, ontologies, corpora) for high-quality cross-lingual/multilingual systems?

It is obvious that different technologies/applications – and different approaches within the same application - need different information types: e.g. the needs of CLIR or content access systems are quite different from MT systems. Do we have examples of really 'good' bilingual/multilingual lexicons, at least for some applications?

Which are the priority information types for different multilingual content management systems? Are we able to establish clear lexical/linguistic/knowledge requirements for different application types, or even component technologies? And to define steps to gradually reach consensus?

[1] http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm.

Which is the respective role of e.g. annotated corpora, monolingual lexicons (with different information types), bi- multilingual lexicons, ontologies, knowledge bases, etc?

Can we aim at basic, general purpose bilingual/multilingual lexicons, to be tuned, adapted to different applications?

A key strategic question - also for the funding agencies - is: for which type of resources to invest? With respect to short vs. medium term results?

Is there the need for robust systems, able to acquire/tune lexical/linguistic knowledge, to accompany static basic resources? in particular, systems able to acquire multilingual lexical/linguistic information? Do we have good sources of bi-/multilingual information (machine readable dictionaries, corpora, …)? And reliable methods for acquisition? Do we have to rely on parallel corpora? Or it is more advisable to aim at the use of 'comparable corpora', accompanied by robust technologies for annotation (at different levels: morphosyntactic, syntactic/functional, semantic, …), and by a shared set of text annotation schemata?

What is the relation between lexical standards and text annotation standards? In particular when we speak about "content" interoperability, is the field 'mature' enough to converge around agreed standards? Or is the market compelling us toward operational standards?

Is the field of multilingual lexical resources ready to tackle the challenges set by the Semantic Web development?

Knowledge management is critical. Is it an achievable goal to arrive at some commonly agreed text annotation protocol also for the semantic/conceptual level (in order to be able to automatically establish links among different languages)?

A last but critical question: if we had real-size lexicons plus conceptual systems with very fine-grained semantic/conceptual information, would there be systems (non ad-hoc toy systems) able to use them? It seems sometimes that there is a loop, or a vicious circle, between i) lack of suitable, large-size and knowledge intensive, resources (lexicons, ontologies, corpora, with many different types of syntactic, semantic, conceptual information encoded), and ii) systems' ability to use them effectively. Should we define a strategy of research and development within which the two paths are pursued in parallel, closely interact with each other, and be gradually integrated?

3. References

Calzolari, N., McNaught, J., Zampolli, A. 1996. "EAGLES Final report: Editors' Introduction", EAG-EB-FR, Pisa.

Calzolari, N., Grishman, R., Palmer, M. (eds.). 2001. "Survey of major approaches towards Bilingual/Multilingual Lexicons". ISLE CLWG Deliverable D2.1-D3.1, Pisa.

Calzolari, N., Lenci, A., Zampolli, A., Bel, N., Villegas, M., Thurmair, G. 2001. "The ISLE in the Ocean. Transatlantic Standards for Multilingual Lexicons (with an Eye to Machine Translation)". In *Proceedings of the MT Summit*, Santiago de Compostela.

Calzolari, N., Zampolli, A., Lenci, A. 2002. "Towards a Standard for a Multilingual Lexical Entry: the EAGLES/ISLE Initiative". In A. Gelbukh (ed.), *CICLing-2002 Third International Conference on Intelligent text processing and Computational Linguistics*, Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg New York.

Sanfilippo, A. *et al.* (1996). *EAGLES Subcategorization Standards.* See http://www.icl.pi.cnr.it/EAGLES96/syntax/syntax.html

Sanfilippo, A., *et al.* (eds.). 1999. *EAGLES Recommendations on Semantic Encoding.* EAGLES LE3-4244 Final Report. See http://www.ilc.pi.cnr.it/EAGLES96/rep2.

Thurmair, G.: OLIF Input Document. (2000). See http://www.olif.net/main.htm

Villegas, M., Bel, N. (2002). "From DTDs to relational dBs. An automatic generation of a lexicographical station out off ISLE guidelines". In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain.