Broadening the Scope of the EAGLES/ISLE Lexical Standardization Initiative

Nicoletta CALZOLARI
Istituto di Linguistica Computazionale, CNR
Area della Ricerca, Via Moruzzi 1
Pisa, Italy, 56100
glottolo@ilc.cnr.it

Alessandro LENCI
Dipartimento di Linguistica, Università di Pisa
Via S. Maria 36
Pisa, Italy, 56100
alessandro.lenci@ilc.cnr.it

Francesca BERTAGNA
Dipartimento di Linguistica, Università di Pisa
Via S. Maria 36
Pisa, Italy, 56100
francesca.bertagna@ilc.cnr.it

Antonio ZAMPOLLI
Istituto di Linguistica Computazionale, CNR
Area della Ricerca, Via Moruzzi 1
Pisa, Italy, 56100
pisa@ilc.cnr.it

Abstract

ISLE is a continuation of the long standing EAGLES initiative and it is supported by EC and NSF under the Human Language Technology (HLT) programme. Its objective is to develop widely agreed and urgently demanded standards and guidelines for infrastructural language resources, tools, and HLT products. EAGLES itself is a well-known trademark and point of reference for HLT projects and products and its previous results have already become *de facto* widely adopted standards. Multilingual computational lexicons, natural interaction and multimodality, and evaluation are the three areas targeted by ISLE. In the first section of the paper we describe the overall goals and methodology of EAGLES/ISLE, in the second section we focus on the work of the Computational Lexicon Working Group, introducing its work strategy and the preliminary guidelines of a standard framework for multilingual computational lexicons, based on a general schema for the "Multilingual ISLE Lexical Entry" (MILE).

1 Introducing EAGLES/ISLE

ISLE (*International Standards for Language Engineering*) is a continuation of the long standing European EAGLES (*Expert Advisory Group for Language Engineering Standards)* initiative (Calzolari *et al.*, 1996), carried out through a number of subsequent projects funded by the European Commission (EC) since 1993. ISLE is an initiative under the Human Language Technology (HLT) programme within the EU-US International Research Co-operation with the aim to develop and promote widely agreed and urgently demanded HLT standards, common guidelines and best practice recommendations for infrastructural language resources (Zampolli, 1998), (Calzolari, 1998), tools that exploit them, and language engineering products.

Object of EAGLES/ISLE are large-scale language resources (such as text corpora, computational lexicons, speech corpora, multimodal resources), means of manipulating such knowledge via computational linguistic formalisms, mark-up languages and various software tools and means of assessing and evaluating resources, tools and products (EAGLES EWG final report, 1996).

EAGLES was set up to determine which aspects of our field are open to short-term *de facto* standardisation and to encourage the development of such standards for the benefit of consumers and producers of language technology, through bringing together representatives of major collaborative European R&D projects, and of HLT industry, in relevant areas. In this respect, more than 150

leading industrial and academic players in the HLT field have actively participated in the definition of this initiative and have lent invaluable support to its execution.

Successful standards are those which respond to commonly perceived needs or aid in overcoming common problems. In terms of offering workable, compromise solutions, they must be based on some solid platform of accepted facts and acceptable practices.

The current ISLE project[1] targets the three areas of :

-multilingual computational lexicons[2],
-natural interaction and multimodality (NIMM)[3],
-evaluation of HLT systems[4].

For *multilingual computational lexicons*, ISLE goals are: i) extending EAGLES work on lexical semantics, necessary to establish inter-language links; ii) designing and proposing standards for multilingual lexicons; iii) developing a prototype tool to implement lexicon guidelines and standards; iv) creating exemplary EAGLES-conformant sample lexicons and tagging exemplary corpora for validation purposes; v) developing standardised evaluation procedures for lexicons.

For *NIMM*, ISLE work is targeted to develop guidelines for: i) the creation of NIMM data resources; ii) interpretative annotation of NIMM data, including spoken dialogue in NIMM contexts; iii) metadata descriptions for large NIMM resources; iv) annotation of discourse phenomena.

For *evaluation*, ISLE is working on: i) quality models for machine translation systems; ii) maintenance of previous guidelines - in an ISO based framework (ISO 9126, ISO 14598).

Three Working Groups, and their sub-groups, carry out the work, according to the EAGLES methodology, with experts from both the EU and US, acting as a catalyst in order to pool concrete results coming from major international/national/industrial projects. Relevant common practices or upcoming standards are being used where appropriate as input to EAGLES/ISLE work. Numerous theories, approaches, and systems are being taken into account as any recommendation for harmonisation must take into account the needs and nature of the different major contemporary approaches.

Results are widely disseminated, after due validation in collaboration with EU and US HLT R&D projects, National projects, and industry.

In the following we concentrate on the Computational Lexicon Working Group (CLWG), trying to describe its specific methodology and its goal of establishing a general and consensual standardized environment for the development and integration of multilingual resources. The general vision adheres to the idea of enhancing the sharing and reusability of multilingual lexical resources, by promoting the definition of a common parlance for the community of multilingual HLT and computational lexicon developers. The CLWG pursues this goal by proposing a general schema for the encoding of multilingual lexical information, the *MILE* (Multilingual ISLE Lexical Entry). This has to be intended as a meta-entry, acting as a common representational layer for multilingual lexical resources.

We describe the preliminary proposals of guidelines for the MILE, highlighting some methodological principles applied in previous EAGLES.

2 The Computational Lexicon Working Group

Existing EAGLES results in the Lexicon and Corpus areas are currently adopted by an impressive number of European - and recently also National – projects and has became the "*de-facto* standard" for LR in Europe. This is a very good measure of the impact – and of the

---

[1] Coordinated by A. Zampolli for EU and M. Palmer for US, see http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm.

[2] EU chair: N. Calzolari; US chairs: M. Palmer and R. Grishman.

[3] EU chair: N. O. Bernsen; US chair: M. Liberman.

[4] EU chair: M. King; US chair: E. Hovy.

resources. With respect to this target, one of the first objectives is to discover and list the (maximal) set of (granular) *basic notions* needed to describe the multilingual level. The *Survey* of existing lexicons (Calzolari, Grishman and Palmer, 2001) has been accompanied by the analysis of the requirements of a few multilingual applications, and by the parallel analysis of typical cross-lingually complex phenomena[5]. The main issue is how to state in the most proper way the translation correspondences among entries in the multilingual lexicon. The passage from source language (SL) to target language (TL) makes it necessary to express very complex and articulated transfer conditions, which have to take into account as difficult and pervasive phenomena as argument switching, multi-word expressions, collocational patterns, etc.

The function of an entry in a multilingual lexicon is to supply enough information to allow the system to identify a distinct sense of a word or phrase in SL, in many different contexts, and reliably associate each context with the most appropriate translation. The first step is to determine, of all the information that can be associated with SL lexical entries, what is the most relevant to a particular task. We decided to focus the work of survey and subsequent recommendations around two major broad categories of application: Machine Translation and Cross-Language Information Retrieval. They have partially different/complementary needs, and can be considered to represent the requirements of other application types. It is necessary in fact to ensure that any guidelines meet the requirements of industrial applications and that they are implementable.

In the Survey, some Korean and Japanese examples were present in the *case study* dedicated to relevant cross-linguistic phenomena, (e. g. sense distinctions according to variation in syntactic frames/semantic type/ domain information, differences in predicate argument structure, argument incorporation, conflation, head switching etc).

2.2 Towards the Recommendation Phase: designing the MILE Architecture

Since the architecture of the PAROLE-SIMPLE lexicons has been selected to provide the necessary bootstrapping basis for the stepwise refinement cycle leading to MILE, we briefly provide here some information about these resources. The design of the SIMPLE lexicons (Bel *et al.*, 2000) complies with the EAGLES Lexicon/Semantics Working Group guidelines (Sanfilippo *et al.*, 1999), and the set of recommended semantic notions.

The SIMPLE lexicons are built as a new layer connected to the PAROLE syntactic layer, and encode structured "semantic types" and semantic (subcategorization) frames. They cover 12 languages (Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish). The common model is designed to facilitate future cross-language linking: they share the same *core ontology* and the same set of *semantic templates*.

The "conceptual core" of the lexicons consists of the basic structured set of "semantic types" (the *SIMPLE ontology*) and the basic set of notions to be encoded for each Semantic Unit (*SemU*): domain information, lexicographic gloss, argument structure, selectional restrictions/preferences on the arguments, event type, links of the arguments to the syntactic subcategorization frames as represented in the PAROLE lexicons, 'qualia' structure, following the Generative Lexicon (Pustejovsky, 1995), semantic relations, etc..

SIMPLE and PAROLE lexicons are layered resources, with links between the morphological and syntactic layers expressed in PAROLE and the semantic information present in SIMPLE.

In its general design, also MILE is envisaged as a highly *modular* and *layered* architecture as described in Calzolari *et al.* (2001). Modularity concerns the "horizontal" MILE organization, in which independent and yet linked modules

---

[5] Contributors are: Atkins, Bel, Bertagna, Bouillon, Calzolari, Dorr, Fellbaum, Grishman, Habash, Lange, Lehmann, Lenci, McCormick, McNaught, Ogonowski, Palmer, Pentheroudakis, Richardson, Thurmair, Vanderwende, Villegas, Vossen, Zampolli.

target different dimensions of lexical entries. On the other hand, at the "vertical" level, a layered organization is necessary to allow for different degrees of granularity of lexical descriptions, so that both "shallow" and "deep" representations of lexical items can be captured. This feature is particularly crucial in order to stay open to the different styles and approaches to the lexicon adopted by existing multilingual systems.

At the top level, MILE includes two main modules, *mono-MILE*, providing monolingual lexical representations, and *multi-MILE*, where multilingual correspondences are defined. With this design choice the ISLE-CLWG intends also to address the particularly complex and yet crucial issue of multilingual resource development through the integration of monolingual computational lexicons. As in the reference model, PAROLE/SIMPLE, Mono-MILE is organized into independent modules, respectively providing *morphological*, *syntactic* and *semantic* descriptions. The latter surely represents the core and the most challenging part of the ISLE-CLWG activities, together with the two other crucial topics of *collocations* and *multi-word expressions*, which have often remained outside astandardization initiatives, and nevertheless have a crucial role at the multilingual level. This bias is motivated by the necessity of providing an answer to the most urgent needs and desiderata of next generation HLT, as also expressed by the industrial partners participating to the project. With respect to the issue of the representation of multi-word expressions in computational lexicons, the ISLE-CLWG is actively cooperating with the NSF sponsored XMELLT project (Calzolari *et al.*, 2002).

Multi-MILE specifies a formal environment for the characterization of multilingual correspondences between lexical items. In particular, source and target lexical entries can be linked by exploiting (possibly combined) aspects of their monolingual descriptions. Moreover, in multi-MILE both syntactic and semantic lexical representations can also be enriched, so as to achieve the granularity of lexical description required to establish proper multilingual correspondences, and which is possibly lacking in the original monolingual lexicons.

According to the ISLE approach, monolingual lexicons can thus be regarded as *pivot lexical repositories*, on top of which various language-to-language multilingual modules can be defined, where lexical correspondences are established by partly exploiting and partly enriching the monolingual descriptions. This architecture guarantees the independence of monolingual descriptions while allowing for the maximum degree of flexibility and consistency in reusing existing monolingual resources to build new bilingual lexicons.

The MILE architecture is intended to provide the common representational environment needed to implement such an approach to multilingual resource development, with the goal of maximizing the reuse, integration and extension of existing monolingual computational lexicons.

In the process of specifying the various components of MILE, the ISLE-CLWG has adopted a two-track strategy:

1) identifying the lexical dimensions and the various types of information which are relevant to establish multilingual correspondences;
2) idefining a suitable formal data model to encode this information as well as the operations required at the multilingual level.

To tackle point 1) the survey of the available computational lexicons (see section 2.1) has been complemented with a more lexicographic-based effort, to identify the types of information used in bilingual dictionaries to establish translation equivalents. To this purpose, the CLWG has organized two "task forces" with the responsibility respectively of creating a sample of lexical entries and investigating the use of the so-called *sense indicators* in traditional bilingual dictionaries. The work on *sense indicators* has been carried out mainly by S. Atkins and P. Bouillon: sense indicators are the 'clues' given by the lexicographer to the bilingual dictionary users in order to guide them to the most appropriate choice of equivalence in the foreign language. The source word with its syntactic category, the target words and the sense indicators were

automatically extracted from an English-French dictionary and then the sense indicators have been classified on the basis of lexical relevant facts (cf. Atkins *et al.*, 2002).
The aim of these activities has been twofold: on one hand, we wanted to be able to highlight the various types of information useful to determine the transfer conditions; on the other, we had to explore and evaluate the full expressive potentialities provided by the reference computational model (i.e. the PAROLE-SIMPLE architecture).

3.2 The MILE Data Structure and Lexicographic Environment

The CLWG is setting up a lexicographic environment consisting of the following four main components: i) the *MILE Entry Skeleton*, ii) the *MILE Lexical Data Categories,* iii) the *MILE Shared Lexical Objects,* iv) the *ISLE Lexicographic Station*.
The *MILE Entry Skeleton*, formalized as an XML DTD, is an Entity Relationship model that will define the general constraints for the construction of multilingual entries, as well as the grammar to build the whole array of lexical elements needed for a given lexical description.
The *MILE Lexical Data Categories* will provide the lexical objects (syntactic and semantic features, semantic relations, syntactic constructions, predicates and arguments etc..) that are the basic components of MILE-conformant lexical entries. Lexical Data Categories will be organized in a hierarchy and will be defined using RDF schema (Brickley and Guha, 2000) to formalize their properties and make their "semantics" explicit.
The *MILE Shared Lexical Objects* will instantiate the MILE Lexical Data Categories, to be used to build in an easy and straightforward way lexical entries. These will include main syntactic constructions, basic operations and conditions to establish multilingual links, macro-semantic objects, such as lexical conceptual templates acting as general constraints for the encoding of semantic units.

For instance, at the multilingual level it is possible to identify a first set of basic operations that are at the basis of multilingual transfer tests and actions. This would include: i) adding to a monolingual lexical entry a new syntactic position (required for a given translation correspondence); ii) adding to a monolingual semantic description a new semantic feature (required for a given translation correspondence); iii) constraining the source-target correspondence to apply only if an existing syntactic position is realized by a certain type of phrase, etc.
Lexical objects will be identified by an URI and will act as a common resources for lexical representation, to be in turn described by RDF metadata. The defined lexical objects will be used by the lexicon (or applications) developers to build and target lexical data at a higher level of abstraction. Thus, they have to be seen as a step in the direction of simplifying and improving the usability of the MILE recommendations.
The ISLE Lexicographic Station is a development platform used to automatically generate a prototype tool starting from the MILE DTD. The aim of this prototype tool is to i) exemplify the MILE entry ii) make extensive use of already existing monolingual resources, and iii) eventually test the guidelines in a real scenario. This situation led us to define a lexicographic station development platform that guarantees the portability of the final prototype to the final specifications as well as to existing monolingual resources which will serve as the basic data for MILE (for a detailed description, cf. Villegas and Bel, 2002).
Both at monolingual and multilingual level (but with particular emphasis on the latter), ISLE intends to start up the incremental definition of a more Object-Oriented layer for lexical description and to foster the vision of open and distributed lexicons, with elements possibly residing in different sites of the web.

3 Enlargement to Asian Languages

An enlargement of the group to involve also Asian languages is going on and representatives of Chinese, Japanese, Korean, and Thai languages have contributed to ISLE work and participated in some ISLE workshops.

The cooperation between Asia and Europe has to be pursued also through new common initiatives, as the expression of interest for the creation of an *Open Distributed Lexical Infrastructure* that has been submitted to the European Commission for the 6th Framework Programme for Research.

This expression of interest is supported by many non-EU participants, as the newly formed Asian Federation of Natural Language Processing Associations (AFNLPA), the Department of Computer Science of the University of Tokyo, the Korean KAIST and KORTERM, the Taiwanese Institute of Linguistics of the Academia Sinica.

The *Open Distributed Lexical Infrastructure,* a natural development of the ISLE model, can be seen as a new paradigm of distributed lexicon creation and maintenance and it would be a step of great importance for the fulfilment of the vision of the Semantic Web (Berners-Lee, 1998). The creation of such infrastructure has to be *consensual* and in this regard needs the collaboration of a group of languages as large as possible (for example the AFNLPA brings into the initiative many Asian languages, such as Chinese, Hindi, Indonesian, Japanese, Korean, Malay, Tamil, Thai and Urdu). A prerequisite in order to reach interoperability is the existence of best practices and standards that have been consensually agreed on or have been submitted to the international community as *de-facto* standards.

4 Conclusions

In this paper we presented overall goals and methodological principles of the standardization activity of EAGLES/ISLE. In particular, we describe the work of the Computational Lexicon Working Group and its effort towards recommendations, focussing on the MILE, the multilingual lexical meta-entry proposed as the standard representational format for multilingual computatio nal lexical resources. Lexical representation is articulated over different information layers, each factoring out different, but possibly inter-related, linguistic facets of information, relevant in order to establish multilingual lexical links. We also pointed out the necessity to involve a broader group of languages in order to ensure the achievement of a real consensual standard.

Acknowledgements

We thank all the members of the ISLE CLWG for their active participation and contribution to this enterprise.

References

Atkins S., Bel N., Bertagna F., Bouillon P., Calzolari N., Fellbaum C., Grishman R., Lenci A., MacLeod C., Palmer M., Thurmair R., Villegas M., Zampolli A. (2002) *From Resources to Applications. Designing The Multilingual ISLE Lexical Entry*. In Proceedings of LREC 2002, Las Palmas, Canary Islands, Spain.

Bel N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli A. (2000) *Simple: A General Framework for the Development of Multilingual Lexicons.* In: LREC Proceedings. Athens.

Berners-Lee T. (1998) *Semantic Web road map*. Personal note. Available at: http://www.w3.org/DesignIssues/Semantic.html.

Burnard L., Baker P., McEnery A., Wilson A. (1997) *An analytic framework for the validation of language corpora*. Report of the Elra Corpus Validation Group.

Calzolari N (1998) An Overview of Written Language Resources in Europe: a few Reflections, Facts, and a Vision. In: Rubio, A., Gallardo, N., Castro, R., Tejada A. (eds.) Proceedings of the First International Conference on Language Resources and Evaluation. Granada 217-224.

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

Calzolari N., Fillmore C.J., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. (2002) Towards Best Practice for Multiword Expressions in Computational Lexicons. In Proceedings of LREC 2002, Las Palmas, Canary Islands, Spain.

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 Machine Translation Summit VIII, Santiago de Compostela, Spain.

Calzolari N., Mc Naught J., Zampolli A (1996) Eagles Final Report: Eagles Editors' Introduction. Pisa.

Eagles (1996) Evaluation of Natural Language Processing Systems. Final Report. CST, Copenhagen. Also at http://issco-www.unige.ch/projects/ewg96/ewg96.html

Brickley D., Guha R. (2000) Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation. Available online at http://www.w3.org/TR/rdf-schema.

Leech G., Wilson A. (1996) Recommendations for the morphosyntactic annotation of corpora. Lancaster.

Lenci A., Busa F., Ruimy N., Gola E., Monachini M., Calzolari N., Zampolli A. (1999) Linguistic Specifications. Simple Deliverable D2.1. ILC and University of Pisa.

Monachini M., Calzolari N. (1996) Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages. ILC-CNR, Pisa.

Monachini M., Calzolari N. (1999) *Standardization in the Lexicon.* In: H. van Halteren (ed.): Syntactic Wordclass Tagging. Kluwer, Dordrecht 149-173.

Pustejovsky J. (1995) *The Generative Lexicon*. Cambridge, MA, MIT Press.

Ruimy N., Corazzari O., Gola E., Spanu A., Calzolari N., Zampolli A. (1998) *The European LE-Parole Project: The Italian Syntactic Lexicon*. In: Proceedings of the First International Conference on Language resources and Evaluation. Granada 241-248.

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

Sanfilippo A. et al. (1999) *Eagles Recommendations on Semantic Encoding*.. See http://www.ilc.pi.cnr.it/EAGLES96/rep2

Underwood N., Navarretta C. (1997) *A Draft Manual for the Validation of Lexica*. Final ELRA Report. Copenhagen.

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.

Zampolli A. (1997) *The PAROLE project in the general context of the European actions for Language Resources*. In: Marcinkeviciene, R., Volz, N. (eds.): Telri Proceedings of the Second European Seminar: Language Applications for a Multilingual Europe. IDS/VDU, Manheim/Kaunas.

Zampolli A. (1998) *Introduction of the General Chairman*. In: Rubio, A., Gallardo, N. Castro, R., Tejada A. (eds.): Proceedings of the First International Conference on Language Resources and Evaluation. Granada, Spain.