International Standards for Multilingual Resource Sharing: The ISLE Computational Lexicon Working Group

Nicoletta Calzolari, Alessandro Lenci, Antonio Zampolli
Istituto di Linguistica Computazionale, CNR, Pisa
Consorzio Pisa Ricerche
Università di Pisa, Dipartimento di Linguistica
[glottolo,lenci,eagles]@ilc.pi.cnr.it

Abstract

The ISLE project is a continuation of the long standing EAGLES initiative, carried out under the Human Language Technology (HLT) programme in collaboration between American and European groups in the framework of the EU-US International Research Co-operation, supported by NSF and EC. We concentrate in this paper on the current position of the ISLE Computational Lexicon Working Group. We provide a short description of the EU SIMPLE lexicons built on the basis of previous EAGLES recommendations. We then point at a few basic methodological principles applied in previous EAGLES phases, and describe a few principles to be followed in the definition of a Multilingual ISLE Lexical Entry (MILE).

1 Introduction: the EAGLES initiative

1.1. What is EAGLES/ISLE?

The ISLE project is a continuation of the long standing EAGLES initiative (Calzolari *et al.*, 1996), carried out through a number of subsequent projects funded by the European Commission (EC) since 1993. EAGLES stands for *Expert Advisory Group for Language Engineering Standards* and was launched within EC Directorate General XIII's Linguistic Research and Engineering (LRE) programme, continued under the Language Engineering (LE) programme, and now under the Human Language Technology (HLT) programme as ISLE, since January 2000. ISLE stands for *International Standards for Language Engineering*, and is carried out in collaboration between American and European groups in the framework of the EU-US International Research Co-operation, supported by NSF and EC. ISLE was built on joint preparatory EU-US work of the previous 2 years towards setting up a transatlantic standards oriented initiative for HLT.

The objective of the project is to support HLT R&D international and national projects, and HLT industry by developing, disseminating and promoting widely agreed and urgently demanded HLT standards and guidelines for infrastructural language resources (see Zampolli, 1998, and Calzolari, 1998), tools that exploit them and LE products. The aim of EAGLES/ISLE is thus to accelerate the provision of standards, common guidelines, best practice recommendations for:

- very large-scale language resources (such as text corpora, computational lexicons, speech corpora (Gibbon *et al.*, 1997), multimodal resources);
- means of manipulating such knowledge, via computational linguistic formalisms, mark-up languages and various software tools;

- means of assessing and evaluating resources, tools and products (EAGLES, 1996).

The basic idea behind EAGLES work is for the group to act as a catalyst in order to pool concrete results coming from current major International/ National/industrial projects. Relevant common practices or upcoming standards are being used where appropriate as input to EAGLES/ISLE work, particularly in the areas of computational lexicons, text, speech, and multimodal annotation, and evaluation. Numerous theories, approaches, and systems are being taken into account, where appropriate, as any recommendation for harmonisation must take into account the needs and nature of the different major contemporary approaches. EAGLES is also drawing strong inspiration from the results of major projects whose results have contributed to advancing our understanding of harmonisation issues.

1.2 A quick Overview of the ISLE Work

The current ISLE project (see http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm) targets the three areas of multilingual computational lexicons, natural interaction and multimodality (NIMM), and evaluation of HLT systems. These areas were chosen not only for their relevance to the HLT call but also for their long-term significance.

- For multilingual computational lexicons, ISLE is working to: extend EAGLES work on lexical semantics, necessary to establish inter-language links; design and propose standards for multilingual lexicons; develop a prototype tool to implement lexicon guidelines and standards; create exemplary EAGLES-conformant sample lexicons and tag exemplary corpora for validation purposes; and develop standardised evaluation procedures for lexicons.

- For NIMM, a rapidly innovating domain urgently requiring early standardisation, ISLE work is targeted to develop guidelines for: the creation of NIMM data resources; interpretative annotation of NIMM data, including spoken dialogue in NIMM contexts; and annotation of discourse phenomena.

- For evaluation, ISLE is working on: quality models for machine translation systems; and 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 already proven EAGLES methodology, with experts from both the EU and US, working and interacting within a strongly co-ordinated framework. International workshops are used as a means of achieving consensus and advancing work. Results will be widely disseminated and published, after due validation in collaboration with EU and US HLT R&D projects, National projects, and industry.

1.3. The Computational Lexicon Working Group

We concentrate in the following on the current position of the ISLE Computational Lexicon Working Group (CLWG).

EAGLES work towards *de facto* standards has already allowed the field of Language Resources to establish broad consensus on key issues for some well-established areas — and will allow similar consensus to be achieved for other important areas through the ISLE project — providing thus a key opportunity for further consolidation and a basis for technological advance. EAGLES previous results have already become *de facto* standards. To mention several key examples: the LE PAROLE/SIMPLE resources (morphological/syntactic/semantic lexicons and corpora for 12 EU languages, Ruimy *et al.*, 1998, Lenci *et al.*, 1999, Bel *et al.*, 2000) rely on EAGLES results (Sanfilippo, A. *et al.*, 1996 and 1999), and are now being enlarged at the national level through many National Projects; the ELRA Validation Manuals for Lexicons (Underwood and Navarretta, 1997) and Corpora (Burnard *et al.*, 1997) are based on EAGLES guidelines; morpho-syntactic tagging of corpora in a very large number of EU, international and national projects – and for more than 20 languages — is conformant to EAGLES recommendations (Leech and Wilson, 1996).

The first priority of the CLWG in the first phase of the ISLE project was to do a comprehensive survey of existing multilingual lexicons. To this end the European and the American members decided, among others, i) to prepare a grid for lexicon description to classify the content and structure of the surveyed resources on the basis of a number of agreed parameters of description, and ii) to provide a list of cross-lingual lexical phenomena that could be used to focus the survey. The inventory (survey) of what exists and is available (semantic and bilingual/multilingual lexicons, printed bilingual dictionaries) is now being completed, and will be made soon available on the Web. Each participant engaged for surveying a number of resources. A list of the main applications that use lexical resources was also established, to focus the survey and subsequent recommendations around them. Each summary of a particular bilingual or multilingual dictionary includes: i) a description of the surveyed dictionary structure (on the basis of the common grid), ii) for one or two examples from the cross-lingual lexical phenomena, an explanation of how these examples are handled by this dictionary.

2 The structure of the prospective Multilingual ISLE Lexical Entry

The main goal of the CLWG is the definition of a Multilingual ISLE Lexical Entry (henceforth MILE). This is the main focus of the second year of the project, the so called "recommendation phase".

2.1 Basic EAGLES principles

We remind here just a few basic methodological principles derived from and applied in previous EAGLES phases. They have proven useful in the process of reaching consensual *de facto* standards in a bottom-up approach and will be at the basis also of ISLE work.

The MILE is envisaged as a highly ***modular*** and possibly ***layered*** structure, with different levels of recommendations. Such an architecture has been proven useful in previous EAGLES work, e.g in the EAGLES morphosyntactic recommendations (Monachini and Calzolari, 1996), which embody three levels of linguistic information: obligatory, recommended and optional (optional splits furthermore into language independent and language dependent). This modularity would enhance: the flexibility of the representation, the easiness of customisation and integration of existing resources (developed under different theoretical frameworks or for different applications), the usability by different systems which are in need of different portions of the encoded data, the compliance with the proposed standards also of partially instantiated entries.

The MILE recommendations should also be very ***granular***, in the sense of reaching a maximal decomposition into the minimal basic information units that reflect the phenomena we are dealing with. This principle was previously recommended and used to allow easier reusability or mappability into different theoretical or system approaches (Heid and McNaught, 1991): small units can be assembled, in different frameworks, according to different (theory/application dependent) generalisation principles. Such basic notions must be established before considering any system-specific generalisations, otherwise our work may be too conditioned by system-specific approaches. For example, 'synonymy' can be taken as a basic notion; however, the notion of 'synset' is a generalisation, closely associated with the WordNet approach. 'Qualia relations' are another example of a generalisation, whereas 'semantic relation' is a basic notion. Modularity is also a means to achieve better granularity.

On the other side, past EAGLES experience has shown it is useful in many cases to accept ***underspecification*** with respect to recommendations for the representation of some phenomenon (and ***hierarchical structure*** of the basic notions, attributes, values, etc.), i) to allow for agreement on a minimal level of specificity especially in cases where we cannot reach wider agreement, and/or ii) enable mappability and comparability of different lexicons, with different granularity, at the minimal common level of specificity (or maximal generality). For example, the work on syntactic subcategorisation in EAGLES proved that it was problematic to reach agreement on a few notions, e.g. it seemed unrealistic to agree on a set of grammatical functions. This led to an

underspecified recommendation, but nevertheless one that was useful.

One of the first objectives of the CLWG will be to discover and list the (maximal) set of (minimal/more granular) ***basic notions*** needed to describe the multilingual level. This task will be facilitated by the survey of existing lexicons, accompanied by the analysis of the requirements of a few multilingual applications, and by the parallel analysis of typical multilingual complex phenomena. Most or part of these basic notions should be already included in previous EAGLES recommendations, and, with different distribution, in the existing and surveyed lexicons. We have therefore to revisit earlier linguistic layers (previous EAGLES work, essentially monolinguistic) to see what we need to change/add or what we can reuse for the multilingual layer. The multilingual layer thus depends on monolingual layers.

2.2 The MILE architecture

The MILE is intended as a *meta-entry*, acting as a common representational layer for multilingual lexical resources. The key-ideas underlying the design of a meta-entry can be summarized as follows. Different theoretical frameworks appear to impose different requirements on how lexical information should be represented. One way of tackling the issue of theoretical compatibility stems from the observation that existing representational frameworks mostly differ in the way pieces of linguistic information are mutually implied, rather than in the intrinsic nature of this information. To give a concrete example, almost all theoretical frameworks claim that lexical items have a complex semantic organization, but some of them try to describe it through a multidimensional internal structure (cf. the *qualia structure* in the Generative Lexicon, Pustejovsky 1995), others by specifying a network of semantic relations (cf. WordNet, Miller *et al.* 1990), and others in terms of argumental frames (cf FrameNet, Baker *et al.* 1998; Lexical Conceptual Structures, Jackendoff 1992; etc.). A way out of this theoretical variation is to augment the expressive power of the lexical representation language both *horizontally*, i.e. by distributing the linguistic information over mutually independent "coding layers", and *vertically*, by further specifying the information conveyed by each such layer. This solution will contribute to solve the issues raised by theoretical variation by defining a common level onto which different types of resources will be mapped without loss of information. This appears to be a necessary condition to guarantee an efficient re-use and interchange of lexical data, often coming from resources developed according to very different architectural and theoretical criteria.

With respect to this issue, the MILE is designed to meet the following desiderata:

- factor out linguistically independent (but possibly correlated) primitive units of lexical information;
- make explicit information which is otherwise only indirectly accessible by NLP systems;
- rely on lexical analysis which have the highest degree of inter-theoretical agreement;
- avoid framework-specific representational solutions.

All these requirements serve the main purpose of making the lexical meta-entry open to task- and system-dependent parameterization.

The MILE is modular along at least three dimensions:

- modularity in the macrostructure and general architecture of MILE
- modularity in the microstructure
- modularity in the specific microstructure of the MILE word sense.

A. ***Modularity in the macrostructure and general architecture of the MILE*** – The following modules should be at least envisaged, referring to the macrostructure of a multilingual system:

1. *Meta-information* - versioning of the lexicon, languages, updates, status, project, origin, etc. (see e.g. OLIF (Thurmair, 2000), GENELEX).
2. Possible architecture(s) of bilingual/ multilingual lexicon(s): we must analyse the interactions of the different modules, and the general structure in which they are inserted, both in the interlingua- and transfer-based approaches, and in possibly hybrid solutions. An open issue is also the relation between the source language (SL) and target language (TL) portions of a lexicon.

B. ***Modularity in the microstructure of the MILE*** – The following modules should be at least envisaged, referring to the global microstructure of MILE:

1. *Monolingual linguistic representation* - this includes the morphosyntactic, syntactic, and semantic information characterizing the MILE in a certain language. It generally corresponds to the typology of information contained in existing lexicons, such as PAROLE-SIMPLE, (Euro)WordNet (EWN), COMLEX, and FrameNet. Following the general organizations of computational lexicons like PAROLE-SIMPLE, which in turn instantiates the GENELEX framework (GENELEX, 1994), at the monolingual level the MILE sorts out the linguistic information into three layers, respectively for morphological, syntactic and semantic dimensions. Typologies of information to be part of this module include (not an exhaustive list):

- **Phonological layer**
  - ◊ phonemic transcription
  - ◊ prosodic information
- **Morphological layer**
  - ◊ Grammatical category
  - ◊ Inflectional class
  - ◊ Modifications of the lemma
  - ◊ Mass/count, 'pluralia tantum'
- **Syntactic layer**
  - ◊ Idiosyncratic behaviour with respect to specific syntactic rules (passivisation, middle, etc.)
  - ◊ Auxiliary
  - ◊ Attributive vs. predicative function, gradability
  - ◊ Subcategorization frames
  - ◊ Grammatical functions of the positions
  - ◊ Morphosyntactic and/or lexical features
  - ◊ Information on control and raising properties
- **Semantic layer**
  - ◊ Characterization of senses through links to an ontology
  - ◊ Domain information
  - ◊ Argument structure, semantic roles, selectional preferences on the arguments
  - ◊ Event type
  - ◊ Link to the syntactic positions
  - ◊ Basic semantic relations between word senses (i.e. synonymy, hyponymy, meronymy)
  - ◊ Description of word-sense in terms of more specific, semantic/world-knowledge relations among word-senses (such as EWN relations, SIMPLE Qualia Structure, FrameNet Frame Elements, etc.)
  - ◊ Information about regular polisemous alternation
  - ◊ Information concerning cross-part of speech relations (e.g. intelligent - intelligence; writer - to write)

The expressive power of the semantic layer is of the utmost importance for the multilingual layer. A general issue discussed in ISLE concerns whether consensus has to be pursued at the generic level of "type" of information or also at the level of its "values" or actual ways of representation. The answer may be different for different notions, e.g. try to reach the more specific level of agreement also on values for types of meronymy, but not for types of ontology.

2. *Collocational information* - This module includes more or less typical and/or fixed syntagmatic patterns including the lexical head defined by the MILE, which can contribute to characterise its use, or to perform more subtle and/or domain specific characterisations. It includes at least:

- Typical collocates
- Support verb construction
- Phraseological or multiwords constructions
- Compounds
- Corpus-driven examples

This module – not yet dealt with in the previous EAGLES - is critical in a multilingual context both to characterise a word-sense in a more granular way and to make it possible to perform a number of operations, such as WSD or translation in a specific context. Here, synergies with the NSF-XMELLT project on multi-word expressions are exploited. First proposals for the representation of support verbs and noun-noun compounds in multilingual computational lexicons are laid out, and now tested on some language pairs.

3. *Multilingual apparatus* – This represents the focal part of the CLWG activities, which will concentrate its main effort in proposing a general framework for the expression of multilingual transfers. Some of the main issues at stake here are:

- identify a typology of the most common cases of problematic transfer (actually this task has been partially performed during the survey phase of the project);
- identify which conditions must be expressible and which transformation actions are necessary, in order to establish the correct multilingual mappings;
- select which types of information these conditions must access in the modules (1) and (2) above;
- identify the various methods of establishing SL --> TL equivalence
- examine the variability of granularity needed when translating in different languages, and the architectural implications of this.

C. ***Modularity in the specific microstructure of the MILE word-sense*** (word-sense is the basic unit at the multilingual level) – Senses should also have a modular structure (i.e. the above distinction between modules (B.1.) and (B.2.) must be intended at word-sense level):

1. *Coarse-grained* (general purpose) characterisation in terms of prototypical properties, captured by the formal means in (B.1.) above, which serves to partition the meaning space in large areas and is sufficient for some NLP tasks.

2. *Fine-grained* (domain or text dependent) characterisation mostly in terms of collocational/syntagmatic properties (B.2.), which is especially useful for specific tasks, such as WSD and translation. Different types of information may have a sort of different operational specialisation.

3 Methodological and organisational issues

As in previous EAGLES, it is considered helpful to base the recommendations on the requirements stemming from a few application systems. The CLWG agreed to focus on two major broad categories of application: machine translation (MT) and cross-lingual information retrieval (CLIR).

As said above, the CLWG has agreed that we should base any multilingual description on monolingual descriptions. MILE should therefore include previous EAGLES recommendations for other layers. We must evaluate the usefulness of these layers with respect to multilingual tasks, focusing in particular on MT and CLIR tasks. Obviously an additional module is needed, where correspondences between languages are defined, including conditions on syntactic structures involving lexical entries. The linking module (transfer) may not be the same for different applications: it may be simpler for CLIR, which may be a subset of the one needed for MT. For CLIR, an ontology or semantic hierarchy is however required.

We are also adopting an approach that would lead to a formalisation of the information contained in traditional bilingual dictionaries, such as restrictions on translation, collocations and examples.

The CLWG agreed the following were appropriate tasks to concentrate on, in order to discover basic notions for MILE:

1. Analyse information given to the human user in bilingual/monolingual dictionaries that allows selection of correct equivalence.
2. Analyse (if these can be obtained) instructions/guidelines supplied to lexicographers for writing bilingual entries.
3. Investigate, in corpus concordances, which are the clues that allow to disambiguate/decide on proper sense for translation.
4. Elaborate a typology of transfer conditions/actions and investigate lexical requirements.
5. Look at multilingual lexical requirements for approaches based on interlingual concepts/ontologies.
6. Rank our typology in terms of scale of difficulty of disambiguation

3.1. Types of information to be addressed

Regarding the various types of information to be addressed, the following "workflow" was agreed:

1. notion already exists in previous work (EAGLES, PAROLE/SIMPLE, EWN, etc.):
   - evaluate the notion to see if it is generally adequate
   - evaluate its usefulness for multilingual purposes
2. notion does not exist as recommendation and is not otherwise used in applications (e.g. collocation type), or there are notions from other layers that we have not already considered:
   - decide which method is needed to do work on it
   - prioritise: what is used already in multilingual lexicons (but not covered in EAGLES, e.g. covered in OLIF) and also then look at what needed in near future
   - record what needs further development.

A starting point will be the previous EAGLES recommendations, as instantiated in PAROLE/SIMPLE, for which – as said above - there is a unique DTD for all the 12 languages involved. This will be revised and augmented after work done on various types of information. ISLE will also implement a lexicographic tool, with which a sample of lexical entries will be encoded according to the MILE structure.

Assignments for in-depth analysis of the information types were done, and work is now carried out by the various CLWG members. Results of on-going work will provide: (i.) a list of types of information that should be encoded in each module; (ii.) linguistic specifications and criteria; (iii.) a format for their representation in multilingual lexicons; (iv.) their respective weight/importance in a multilingual lexicon (towards a layered approach to recommendations).

4 Conclusions

Lexicon construction is a costly enterprise, and a major goal is to set up general initiatives to ease and optimise this process. The crescent needs of lexical data, both of general and of domain-specific nature, makes lexicon development an always incremental and potentially open effort, often to be carried out in distributed environments and through the joint work of multiple actors. It is therefore necessary to facilitate lexicon versioning and authoring, the fast integration and scalability of the resources, the fast integration of domain and general linguistic knowledge, as well as the integration of the work of human lexicographers with the information automatically extracted from corpora and dictionaries. The main purpose of the ISLE CLWG is to provide a satisfactory answer to these needs, by establishing a general infrastructure for lexical resources sharing. Its backbone is represented by the MILE, a lexical meta-entry, whose definition is now the focus of the CLWG activities. The MILE is a modular architecture for the representation of multilingual lexical data, and aims at becoming a common parlance for the representation and encoding of lexical data.

References

Baker, Collin F., Fillmore, Charles J., and Lowe, John B. (1998). "The Berkeley FrameNet project"; in *Proceedings of the COLING-ACL*, Montreal, Canada.

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. *LREC Proceedings*, Athens.

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 A. Rubio, N. Gallardo, R. Castro, A. Tejada (eds.), *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, pp.217-224.

Calzolari, N., Mc Naught, J., Zampolli, A. (1996). *EAGLES Final Report: EAGLES Editors' Introduction*. EAG-EB-EI, Pisa.

EAGLES (1996). *Evaluation of Natural Language Processing Systems.* Final Report, Center for Sprogteknologi, Copenhagen. Also available at http://issco-www.unige.ch/projects/ewg96/ewg96.html.

GENELEX Consortium, (1994). *Report on the Semantic Layer*, Project EUREKA GENELEX, Version 2.1.

Gibbon, D., Moore R., Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, New York.

Heid, U., McNaught, J. (1991). *EUROTRA-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications*. Final report.

Jackendoff, R. (1992), *Semantic Structures*, Cambridge, MA, MIT Press.

Leech, G., Wilson, A. (1996). *Recommendations for the morphosyntactic annotation of corpora*, Eag-tcwg-mac/r, ILC-CNR, Pisa.

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.

Miller G.A, Beckwidth R., Fellbaum C., Gross D., and Miller K.J. (1990), "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography*, III, No.4: 235-244.

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

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

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

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

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