Introduction to the Final CLWG Guidelines

Deliverable D1.1
ISLE Computational Lexicon Working Group

Responsible Authors

Nicoletta Calzolari Istituto di Linguistica Computazionale, CNR, Pisa, ITALY
Consorzio Pisa Ricerche, Pisa, ITALY
Ralph Grishman New York University, New York City, NY, USA
Martha Palmer CIS Department, University of Pennsylvania, Philadelphia, PA, USA

Authors

Francesca Bertagna[1 & 2], *Alessandro Lenci*[2], *Monica Monachini*[1]

1: Istituto di Linguistica Computazionale, CNR, Pisa, ITALY
2: Università di Pisa, Pisa, ITALY

1 The ISLE Computational Lexicon Working Group 3
1.1.1 Survey of major approaches to Bilingual/Multilingual Lexicons 5
(Deliverable D2.1-D3.1) 5
1.1.2 Standards and best practice for multilingual computational lexicons 6
(Deliverable 2.2&3.2) 6
1.1.3 Lexical Entry Tool (software) and Accompanying Documentation Prototype 7
1.1.4 Creating MILEs: 100 lexical entries for 3 languages cross-linked (Del. D5.1) 7
1.1.5 ISLE Deliverable 5.2: Sense Tag Annotation of Multilingual Corpora 8
(Deliverable D5.2) 8
1.1.6 Testing Multilinguality in MILE (Deliverable D6.1) 8
1.1.7 MILE Potential Users' Evaluation and Feedback (Deliverable D6.2) 8

1 The ISLE Computational Lexicon Working Group

ISLE[1] (*International Standards for Language Engineering*), a transatlantic standards oriented initiative under the Human Language Technology (HLT) programme, is a continuation of the long-standing European EAGLES initiative (Calzolari, Mc Naught and Zampolli, 1996), carried out through a number of subsequent projects funded by the European Commission (EC) since 1993 (coordinated by A. Zampolli for the Consorzio Pisa Ricerche). 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 under the Human Language Technology (HLT) programme as ISLE, since January 2000. ISLE is carried out by European and American groups within the EU-US International Research Co-operation, supported by EC and NSF. ISLE was built on joint preparatory EU-US work of the previous two years aimed at 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 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 ISLE project (coordinated by A. Zampolli for EU and M. Palmer for US) (see http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm) targets the three areas of

- *multilingual computational lexicon* (EU chair: N. Calzolari; US chairs: M. Palmer and R. Grishman),
- *natural interaction and multimodality* (*NIMM*) (EU chair: N. O. Bernsen; US chair: M. Liberman),
- *evaluation of HLT systems* (EU chair: M. King; US chair: E. Hovy).

These areas were chosen not only for their relevance to the HLT field, but also for their long-term significance. Three Working Groups, and their sub-groups, have carried 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. Responsible partners recruit members from the HLT community (from both academia and industry) to participate in working groups.
International workshops are used as a means of achieving consensus and advancing work. Results are widely disseminated, after due validation in collaboration with EU and US HLT R&D projects, National projects, and industry.

The **ISLE Computational Lexicon Working Group (CLWG)** is committed to the consensual definition of a standardised infrastructure to develop multilingual resources for HLT applications, with particular attention to the needs of Machine Translation (MT) and Crosslingual Information Retrieval (CLIR) systems. Compared with other standardisation initiatives active in this field (e.g. OLIF-2), the original character of ISLE resides in its specifically focusing on the *gray area* of HLT where well-assessed language technology meets more advanced levels and forms of linguistic description. The ISLE CLWG aims at selecting mature areas and results in computational lexical semantics and in multilingual lexicons, which can also be regarded as established achievements, thus to be used as the basis for future research.

[1] ISLE Web Site URL: lingue.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

For *multilingual computational lexicons*, ISLE objectives, as stated in the Technical Annex, are: 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.

In particular, the ISLE-CLWG pursues this goal by designing MILE (*Multilingual ISLE Lexical Entry*), a general schema for the encoding of multilingual lexical information. This has to be intended as a meta-entry, acting as a common representational layer for multilingual lexical resources. Obviously ***MILE also includes previous EAGLES recommendations for other layers***.

Finally, one of the targets of standardisation, and actually one of the main aims of the CLWG activities, is to create a common parlance among the various actors (both of the scientific and of the industrial R&D community) not only in the field of computational lexical semantics and multilingual lexicons, *but also in the areas e.g. of ontologies and the emerging semantic web*, so that synergies will be enhanced, commonalties strengthened, resources and findings usefully shared. In other terms, the process of standard definition undertaken by CLWG, and by the ISLE enterprise in general, on the one side represents an essential interface between advanced research in the field of multilingual lexical semantics, and the practical task of developing resources for HLT systems and applications. It is through this interface that the crucial trade-off between research practice and applicative needs will actually be achieved. On the other side ***ISLE results pave the way to a needed cooperation between until now separate communities, such as HLT actors and groups specifically involved with 'content' (ontologies, semantic web, content providers, etc.)***, enabling future common efforts and resource sharing.

One of the first objectives of the CLWG was to discover and list the (maximal) set of (granular) *basic notions* needed to describe the multilingual level. Since a substantial part of the basic notions should be already included in previous EAGLES recommendations, and, with different distribution, in the existing and surveyed lexicons, and since the multilingual layer depends on monolingual layers, we had to revisit earlier linguistic analysis (previous EAGLES work, essentially monolinguistic) to see what we need to change/add or what we can reuse for the multilingual layer. *Sense distinctions are especially important for multilingual lexicons, since it is at this level that cross-language links need to be established.* The same is true of syntagmatic/collocational/ contextual information. To these areas we have paid particular attention in the recommendation phase, and we have examined how to extend the available EAGLES guidelines in these and other areas to propose a broad format for multilingual lexical entries which is of general usefulness to the community.

The principle guiding the elicitation and proposal of MILE basic notions in the recommendation phase has been, according to a previous EAGLES methodology, the so-called *'edited union'* (term put forward by Gerald Gazdar in earlier EAGLES work) of what exists in major lexicons/models/dictionaries, at least as a starting point, enriched with those types of information which are usually not handled, e.g. those of collocational/syntagmatic nature, and obviously those pertinent to the multilingual layer. This method of work has proven useful in the process of reaching consensual *de facto* standards in a bottom-up approach and is also at the basis of ISLE work. There is a great interest in building on existing resources, rather than starting from scratch; thus efforts must continue in this direction.

In its general design, MILE is envisaged as a highly *modular* and *layered model*. Modularity concerns the "horizontal" MILE organisation, in which independent and yet linked modules target

different dimensions of lexical entries. On the other hand, at the "vertical" level, a layered organisation 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:

- Entity Relationship model
- Repository of shared lexical objects
- Towards sharable and distributed lexical resources (for the semantic web)

This way, both at the monolingual and at the 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. 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. Therefor, lexical objects have to be seen as a step in the direction of simplifying and improving the usability of the MILE recommendations.

Not only will computational lexicons contribute to the content-based management of information on the Web, but also the tools and resources that are being developed for the Semantic Web provide the ground for the architecture and design of next-generation language resources. Moreover, computational lexicons should be conceived as *dynamic systems*, whose development needs to be complemented with the automatic acquisition of semantic information from texts. Gaining insights into the deep interrelation between representation and acquisition issues is likely to have significant repercussions for the way linguistic resources will be designed, developed and used for applications in the years to come. As the two aspects of knowledge representation and acquisition are profoundly interrelated, progress on both fronts can only be achieved, in our view of things, through a full appreciation of this deep interdependency.

The work of the CLWG is articulated on seven deliverables, dedicated to the correspondent Work Packages.
In what follows, we briefly illustrate the deliverables and their content.

1.1.1 Survey of major approaches to Bilingual/Multilingual Lexicons (Deliverable D2.1-D3.1)

The ISLE Deliverable D2.1-D3.1 from the ISLE Computational Lexicons Working Group is based on previous EAGLES work on monolingual lexicon standards, and presents a survey of the major existing bilingual and multilingual lexical resources. Each resource has been investigated to determine its lexical structure and how it encodes cross-language relationships. The survey documents the results in a format that allows easy comparison of the lexical mechanisms employed by each lexicon considered.
The emphasis of the survey is on the semantic level of description as benefits a multilingual perspective, however other levels of lexical information are also taken into account, where these have a bearing on interpretation and on cross-language mapping. This work was undertaken to enable the subsequent identification of a set of basic notions needed to describe the multilingual level, together with other notions that may be recommended for particular purposes or languages.

1.1.2 Standards and best practice for multilingual computational lexicons (Deliverable 2.2&3.2)

The deliverable is the result of the activity of two work packages devoted, on one hand, to the extension of the previous EAGLES work on Lexical Semantics and on the other hand to the actual recommendations and guidelines for the representation of information in bi-multilingual lexica.
The deliverable can be conceived as a two-part document: the first *macrosection* (covering the chapter 1 to 5) concerns the general description of the project and the "background" activity devoted to linguistic/lexicographic analysis aimed at the identification of types of information that play a crucial role in a bi-multilingual environment.
In particular, **Chapter 3** describes EAGLES/ISLE CLWG specific methodology and its goal of establishing a general and consensual standardized environment for the development and integration of multilingual resources.
Chapter 4 briefly presents the first phase of activities of the CLWG, dedicated to the elaboration of a survey of existing multilingual resources both in the European, American and (although still in a more limited extension) Asian research and industrial scenarios. Such a review is also the basis for the process of standard selection and definition, which was the focus of the second phase of the CLWG, aiming at individuating hot areas in the domain of multilingual lexical resources, which call – and *de facto* can access to – a process of standardisation. A specific deliverable (namely the D2.1&3.1) is dedicated to the survey phase. In Chapter 4 a brief description of the two major background lexical resources is also provided: PAROLE/SIMPLE semantic lexicons built on the basis of previous EAGLES recommendations, and WordNet lexicons.
Chapter 5 is dedicated to the results of two different (yet interlinked) lexicographic studies leading to the identification of the information that plays a role, at any level of linguistic description, in the multilingual transfer: section 5.1 deals with a concrete entry creation activity (about 100 multilingual entries for French, Italian and English), section 5.2 describes the use of the so-called *sense indicators* as candidates for transfer conditions. To the work of entry creation a specific deliverable has been dedicated (the D5.1).
The second macrosection goes from **Chapter 6** onwards, where we describe the ISLE proposals of guidelines for the "Multilingual ISLE Lexical Entry" (MILE), the general architecture and features of MILE, as well as discuss the methodology adopted for its definition. The general architecture of MILE is presented in section 6.1. Then, we focus on the two essential ingredients for the MILE specification: on the one hand, the selection of the types of lexical information most relevant to establish multilingual correspondences, the MILE basic notions (section 6.2), and, on the other hand, the specification of the MILE linguistic data model, which will provide the formal backbone of the MILE as a general representation language to develop multilingual resources and to link computational lexicons (6.3). The formalisation of the MILE in RDF is given in section 6.4. MILE can be considered as a platform for a "new generation" of lexical and knowledge resources in the framework of an open and distributed lexical infrastructure.
The description of the MILE prototype tool to implement lexicon guidelines and standards can be found in **Chapter 7.** The tool is a lexicographical station development platform in order to automatically map the DTD into a relational dB, build up a user-friendly interface able to cover the most common requirements of a lexicographic station and to exemplify, test and validate the goodness of the MILE model in a real scenario, that is, reusing already existing monolingual resources such as PAROLE and SIMPLE lexicons.

Other important sections of the deliverable are the appendices, dedicated to encoding examples, case studies, examplificative data etc.
Appendix A, B and C are dedicated to RDF formalization of the MILE: the first appendix present an RDF Schema for MILE, appendix B exemplifies the syntactic layer for some lexical entries and Appendix C defines lexical objects of the *Lexical Data Category Registry* (LDCR).
In Appendices D and E two important reference ontologies are presented, respectively the one from SIMPLE and the EuroWordNet Top Ontology.
One of the important issues ISLE has to deal with is the problem of the representation of noun compounds and support verbs in MILE: a study on the subject, conducted in collaboration with the XMELLT project, is presented in Appendix F.
A crucial aspect in establishing a real and broad consensus is played by communication and sharing of information among many groups active in the field. For this reason we involved also Asian collegues in the ISLE initiative, and we have explored ways of establishing formal links with them. Appendix G provides an exemplification of this collaboration, showing how the EAGLES/ISLE basic notions have been "tested" in the representation of some Asian languages.
Another important issue is the broadening of the scope of ISLE towards a more general and comprehensive description and representation of the linguistic "universe": in Appendix H a survey of the things that are distinctive of the Spoken in respect of Written Language is presented. This can considered a first, important step towards the integration of the ISLE Basic Notions with information important for "wider" and multimodal aspects.

1.1.3 Lexical Entry Tool (software) and Accompanying Documentation Prototype (Deliverable D4.1)

The ISLE deliverable D4.1 describes (i) the lexicographic station development platform used to automatically generate a prototype tool out of theISLE guidelines formally implemented in a DTD, (ii) an actual implementation of the ISLE guidelines expressed in MILE, (iii) the use of the lexicographic station development platform for generating a prototype lexical tool for MILE/ISLE guidelines and the functionalities and (iv) the user manual of the resulting prototype lexicographical station.
The central outcome of ISLE is to define a general schema for multilingual lexical entry (MILE) that is to be the basis for a standard framework for multilingual computational lexicons. In addition, ISLE is to develop a prototype tool to assist the development of multilingual lexical resources in MILE format.
The aim of this prototype tool is to:

- exemplify and disseminate the MILE entry using an actual model with already existing monolingual data,
- assist the development of multilingual lexical resources following MILE schema,
- make extensive use of already existing PAROLE and SIMPLE lexicons and
- to eventually test the effectiveness of the guidelines by using a real scenario.

1.1.4 Creating MILEs: 100 lexical entries for 3 languages cross-linked (Del. D5.1)

The ISLE Deliverable D5.1 describes the task of "creating sample" of MILE-conformant entries. The task has twofold motivations: on the one hand, at the beginning of the project, a case-study phase devoted to a lexicographic analysis was necessary, whose results have been of input for the recommendations. On the other hand, as the MILE design was established, the entry creation has

been important to "strech" the potentialities of the model and to verify whether it was sufficiently flexible so as to permit the solving of lexicographic or translation problems.
The deliverable is organized in: i) a first section describing the "case study" phase and ii) a second part introducing the "testing" of the model in a real entry-encoding scenario.
The deliverable is accompanied by the set of encoded entries for the Italian-to-English and French-to-Italian pairs.

1.1.5 ISLE Deliverable 5.2: Sense Tag Annotation of Multilingual Corpora (Deliverable D5.2)

ISLE Workpackage 5, for the Computational Lexicon Working Group, includes Task 5.2: the annotation of a sample corpus with sense tags. The execution of this task was carried out in concert with SENSEVAL-2: The Second International Workshop on Evaluating Word Sense Disambiguation Systems, which was held on July 5-6, 2001. This paper gives an overview of the creation of the sense tagged data, discussing the evaluation exercise, the tasks, the scoring system, the results, and information about the availability of the annotated corpora. A companion paper is appended, Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically, which has been accepted by the Journal of Natural Language Engineering and is currently under revision. The authors are Martha Palmer, Hoa Dang and Christiane Fellbaum, and this paper addresses the issue of sense granularity, one of the topics for Deliverable 5.2.

1.1.6 Testing Multilinguality in MILE (Deliverable D6.1)

In the Deliverable 6.1 we consider to what extent the multilingual aspects of two different machine translation (MT) systems can be expressed in MILE, the standard proposed in the ISLE project for representing multilingual aspects of lexical entries. The two MT systems are the interlingual system Rosetta, and the transfer system Globalink.

1.1.7 MILE Potential Users' Evaluation and Feedback (Deliverable D6.2)

The evaluation activities pertaining to WP6 start with the desing of an heterogeneous group of potential users of the MILE. In this direction, experts from the lexicon area and an industrial component were involved, in order to ensure business relevance and market impact. Other candidates are partners of other EC and US funded projects dealing with multilingual issues, as well as vendors of language resources such as ELRA and the LDC.
EAGLES/ISLE, which represents a consolidated tradition for western languages, aimed at facing a larger linguistic universe and extending the cooperation world-wide, thus involving languages with other tradition, such as Asian ones.
In addition, a strong potential user is considered to be the newly formed Co-ordination Group of National Projects on Language Resources in Europe, which currently represents 13 European languages, the ENABLER Network.
This group has an obvious interest in assessing early EAGLES results because i) one of its priorities has been declared to be 'Multilingual Language Resources' and ii) it is the developer and owner of the largest set of language resources in Europe.

All the material (the early recommendations, the partial proposals, discussion papers, etc.) was readily put in the public domain on the ISLE CLWG Web Site, in order to allow the potential users to have an easy access to the recommendations.

A crucial aspect in establishing a real and broad consensus is played by the creation of meeting opportunities for discussing, collecting feedback, sharing of information among many groups active

in the field. The major event where it has been possible to evaluate recommendations within ISLE community and representatives of the user group was the Final CLWG Workshop. In the Deliverable 6.2 some of the Workshop results are exposed, together a concrete proposal for further initiatives and collaborations aiming at making MILE a common parlance between different existing lexicons/knowledge bases and models.

1.2 Final Remarks

The CLWG is exploring the notion of new-generation, open, distributed and shared lexical resources. It recognises that a close relationship exists between computational lexicons, content processing, ontologies and the Semantic Web, in that all are concerned not just with format and encoding, but also with content, and content expressed in many languages. This close relationship is however implicit rather than explicit: the CLWG has taken a first step, at its end of project workshop, to involve the other domains mentioned above in discussions in order to emphasise the importance and central role of computational lexicons. MILE is viewed as the beginning of a common platform on which reusable, open, distributed and shared lexical resources may be built that will allow interoperability of content. However, major efforts are implied here, and thus a suitable infrastructure is required. An expression of interest entitled "European Lexical Infrastructure and Technology" (ELITE) has been submitted to the EC within the 6th FP with a view to pursuing this vision in the future. In the meantime, EAGLES will continue its work, coordinating efforts to, for example, further evaluate MILE to determine how easy it is to map different lexicon models to MILE. Mappings to and from OLIF resources, (Euro)WordNet, PAROLE/SIMPLE, ComLex, FrameNet lexicons and others will be examined in this respect, as participants in the end of project workshop undertook to become involved in helping to test MILE. Initial future work will be focussed on defining a methodology to allow such widespread testing to be undertaken, which will also help to raise the credibility of the ELITE initiative. Further future work is planned to deepen analysis of sense-tagging results from the ISLE/SENSEVAL-2 initiative.

Importantly, also, future work will consolidate and extend cooperation with Asian countries. ISLE guidelines will be extended to deal with the many particularities of the universe represented by Asian languages. It is highly noteworthy that the ELITE initiative was strongly supported by many non-EU entities, such as the newly formed Asian Federation of Natural Language Processing Associations (AFNLPA), the Academica Sinica and KORTERM (Korea). Within the EU, the newly formed Coordination Group of National Projects on Language Resources in Europe (ENABLER, currently representing 13 Eueopean languages) has declared a deep interest in assessing and implementing EAGLES results. Thus, experiments with mapping between existing lexicons and MILE, and growing cooperation with key bodies and initiatives on a world-wide basis, will pave the way towards the creation of an open, distributed lexical infrastructure (ODLI), which will represent a new paradigm of distributed lexicon creation and maintenance, that will strongly support the full realisation of the Semantic Web. The work of EAGLES CLWG towards elaboration of consensual standards and best practice thus continues in the foreseeable future with a definite direction and an ever-widening circle of interested and motivated participants.