MILE Users' Evaluation and Feedback

Deliverable D6.2
ISLE Computational Lexicon Working Group

Authors

Nicoletta Calzolari[1 & 2], Francesca Bertagna[1 & 3],
Alessandro Lenci[3], Monica Monachini[1]

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

1.1 Using Feedback from Users as External Evaluation

An important factor in carrying out the evaluation task is the design of a heterogeneous "User Group".
In this perspective, during the activities pertaining to WP6, besides experts from the lexicon area, a industrial component was involved, in order to ensure business relevance and market impact. Moreover, industrial users help establish criteria for measuring performance improvement that can provide continued assessment of EAGLES results.
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.
One of the main achievements of the project was the involvement of Asian colleagues in view of a future world-wide co-operation.
At CLWG Final Workshops in Pisa, representatives of Chinese, Japanese, Korean, Thai and other Asian languages were present. The group has contributed to ISLE work and participated actively to discussions. Asian colleagues can be regarded as "very special" users of MILE, given the very nature of their languages, that led them to concentrate their analysis on:

i) general issues concerning the lexicon, its internal organization, the definition of its atoms, i.e. what a lexical entry is, in particular, from the point of view Asian language typology;
ii) from a more applicative point of view, the results of the application of the EAGLES recommendations focussing primarily on morphological/morphosyntactic aspects (which were the object of the previous EAGLES phase).

The concrete results of this effort to apply the EAGLES/ISLE Guidelines to Asian languages are two documents by Chu-Ren Huang and Tokunaga Takenobu which have been inserted in a specific Appendix of the Deliverable3.1&3.2 (The ISLE Recommendations).

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, chaired by J. Tsujii), the Department of Computer Science of the University of Tokyo, the Korean KAIST and KORTERM, the Taiwanese Institute of Linguistics of the Academia Sinica.

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', ii) it is the developer and owner of the largest set of language resources in Europe, iii) it is strongly interested in ISLE Standardization effort for the repercussions it has in the design of the BLARK (Basic Language Resource Kit).

All the material (the early recommendations, the partial proposals, discussion papers, etc.) was readily put in the public domain in the ISLE CLWG Web Site, in order to allow the potential users to have 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 held in Pisa in December 2002.

1.2 ISLE/EAGLES Workshop: MILE (Multilingual ISLE Lexical Entry) as a Step towards Sharable Multilingual Resources

The primary aim of the ISLE CLWG Workshop was the presentation of the Working Group results. The purpose was not only to circulate the ISLE proposals within the members of the lexicon area, but also to involve neighbouring communities (i.e. Ontologies, Semantic Web, Industrial Multilingual Technologies), obtaining a first evaluation and feedback from them. There was a great attendance of various representatives from the academic and industrial world, who have ensured different typologies of users be mirrored.

The report is organized in two parts:

1. in the first one, an outline of the ISLE workshop and the presentations given by the invited speakers will be provided.
2. in the second part, the discussion stimulated by the panelists during the ending session (the panel) regarding ISLE themes and future initiatives will be outlined.

1.2.1 Section 1: Workshop program and presentations

1. Introduction by the CLWG Co-oordinator

Nicoletta Calzolari opens the session with a presentation of the aims of the Computational Lexicon WG and of the workshop.

The work of the CLWG has been devoted to: i) provide a common environment for development, integration, interchange & sharing of lexical resources with various types of linguistic information, ii) establish a virtuous circle betw. research, applications, & standardization process: lay down a bridge betw. the worlds of research & application iii) mark the boundary between well-consolidated practice and theoretical achievements in multilingual HLT, and areas still open to research but critical for future technological improvements.

A first presentation of the MILE (Multilingual ISLE Lexical Entry), its main features and work methodology was given, together with a statement about the need of broadening the scope of ISLE to other languages and other communities.

All the reference documents can be downloaded at the following address: http://lingue.ilc.pi.cnr.it/EAGLES96/isle/presentations.htm.

Reference document (slides): "Standards and Best Practice for Multilingual Computational Lexicons" (zipped PPT format) by Nicoletta Calzolari (CPR & ILC-CNR)

2. Towards ISLE Recommendations

The session was dedicated to the presentation of the work beyond the design of the MILE: the identification of the MILE Basic Notions.

a. Atkins & Bouillon present an investigating of the use of *sense indicators* in traditional bilingual dictionaries and the tool to handle them. Reference document (slides): "The Lexicographic Approach" (zipped PPT format) by Sue Atkins (WTC) and Pierrette Bouillon (ISSCO)

b. Bertagna & Monachini present the work devoted to the identification of basic lexical dimensions & info-types relevant to establish multilingual links and typology of lexical multilingual correspondences (relevant conditions & actions), carried out through the creation of sample multilingual lexical entries and through the analysis of previous EAGLES results.
Reference document (slides): "From Lexical Entries to MILE Basic Notions" (zipped PPT format) by Francesca Bertagna (ILC-CNR & UniPi) and Monica Monachini (ILC-CNR)

3. *The MILE (Multilingual ISLE Lexical Entry) Model*

a. Lenci & Ide presents the formalization of the ISLE Basic Notions in the actual MILE Lexical Model (MLM), together with a concrete instantiation of the syntactic layer in RDF.

The key issues in the desing of the MILE are: Modularity, User-adaptability, Resource sharing, Reusability. The kernel notions of the MLM are i) the design as an E-R model ii) the lexical objects and the ways they can be combined into a lexical entry.

For further details see the reference document (slides): "The MILE Lexical Model: Linguistic and Formal Architecture" (zipped PPT format) by Alessandro Lenci (UniPi & ILC-CNR) and Nancy Ide (Vassar College)

b. Villegas and Bel present a demo of the functionalities of the tool to manage the MILE. The tool can be viewed as a Lexicographic Station Development Platform, allowing the user to encode, load, dump and browse lexical entries taking in input any ER-model & DTDs.

Reference documents (slides): "ISLE Lexical Tool" (zipped PPS format) by Marta Villegas (gilcUB) and Núria Bel (gilcUB).

4. *Multilingual Technology*

Two presentations regarding industrial applications managing multilingual content were given.

a. Thurmair presents the Concept Manager, a tool for handling resources for CLIR (query expansion & translation). Reference document (slides): "ConceptManager: a Tool to Administer Multilingual ConceptNets" (zipped PDF format) by Gregor Thurmair (Comprendium) and Others (Comprendium).

b. Suzuki presents the MT application developed at Microsoft, based on a hybrid of symbolic and statistical techniques. Reference document (slides): "MSR-MT: a Multilingual Translation System" (zipped PPT format) by Hisami Suzuki (Microsoft Research)

5. *More about ISLE: Spoken Language, Evaluation and Feedback*

a. An important issue that ISLE has to consider is the possibility to represent notions needed for Spoken Language resources. In the future, an experiment for providing a mapping between some of the notions of the Spoken Language and the MILE Basic Notions for written resources will be carried out. Gibbon presents his report: Reference Document (doc): "Resources for Spoken Language and Multimodal Lexica in Multilingual Contexts" (zipped PDF format) by Dafydd Gibbon (UniBielefeld).

b. Peters presents the ISLE Metadata Initiative regarding lexicons. Reference Document (slides): "Lexical Metadata" (zipped PPT format) by Wim Peters (UniSheffield)

c. Reeves presents an experiment of entry creation using MILE, discussing some crucial issues about the representation of collocational and distributional aspects. Reference document (slides): "Using the MILE Model" (zipped PPT format) by Ruth Reeves (NYU), Alessandra Trecci (NYU) and Ralph Grishman (NYU)

d. Odijk's presentation concerns the ways in which multilingual content and multiwords are dealt with in the Rosetta and Globalink projects, and makes proposals wrt MILE. Reference Document (slides): "Multilinguality, MWEs and MILE" (zipped PPT format). Jan Odijk (UniUtrecht).

6. *Feedback on the EAGLES/ISLE Proposal from Asian Languages*

The three slots are dedicated to the Asian perspective on MILE, focussing on the requirements of Asian languages for the lexicon creation.

a. Tokunaga stresses the information needed for the representation of Asian language at morphosyntactic level (taking as input the EAGLES document on MoSy). Reference document (slides): "Feedback on the EAGLES/ISLE Proposal from Asian Languages" (zipped PDF format) by Takenobu Tokunaga (TIT)

b. Huang tackles issues concerning the structure of lexicon with special attention to the orthography (a very important topic for the representation of Asian languages in multilingual perspective). Reference document (slides): "Issues in the Structure of Lexicons: a Multilingual Perspective" (zipped PPT format) by Chu-Ren Huang (Academia Sinica).

c. Finally, Isahara provides an overview of the available resources at CRL, Japan. Reference document (slides): "Linguistic Resources compiled at the Communications Research Laboratory, Japan" (zipped PPT format) by Hitoshi Isahara (CRL)

7. *Open Lexicons, Ontologies and the Semantic Web*

The seventh session host three presentations concerning the emerging field of semantic web and ontologies.

a. Pease provides an overview of the structure of SUMO, an ontology which incorporates 50 sources of high-level ontological content and illustrates its uses in linguistics. Reference document (slides): "The Suggested Upper Merged Ontology (SUMO) and its Use in Linguistics" (zipped PPT format) by Adam Pease (Teknowledge).

b. Gómez-Pérez introduces the notion of Ontology and its potential uses in the general framework of the semantic web, focussing in particular on the Esperonto and OntoWeb experience. Reference

document (slides): "An Insight into Ontologies, Semantic Web and NLP" (zipped PDF format) by Asunción Gómez-Pérez (UPM)

c. Declerck and Buitelaar present a possible approach to Web-based standards lexical semantics and also some ongoing results in three EU projects at DFKI and at the University of the Saarland. Reference document (slides): "Linguistic Resources for Semantic Webs" (zipped PPT format) by Thierry Declerck (UniSaarland) and Paul Buitelaar (DFKI).

1.2.2 Section 2: Panel Session, Discussion and Future Actions

Nicoletta Calzolari opens the discussion provoking the participants with some topics concerning the future of ISLE and, more generally, emerging scenarios for the lexicon community.

ISLE deals with standards for Lexical Resources and multilinguality, with a special interest to standards for content management. The requirements of content processing technologies are at the very base of the Semantic Web enterprise and both the adjacent communities see the importance of reaching consensus on (linguistic & non-linguistic) "content", in addition to agreement on formats & encoding issues. But, is the field of multilingual lexical resources ready to tackle the challenges set by the Semantic Web development?

We should try to establish a continuum line between NLP, lexicons, content, ontologies and Semantic Web in order to reach "content" interoperability. The panelists' and participants' statements show a raising need for an infrastructure for constructing and sharing language resources.
The "vision" of such an infrastructure opens the way towards a "new generation" of LRs: *open*, *distributed* and *shared*.
Moreover, there is the common need to go from *static* (encoded by hand, with methodologies typical of the lexicographic tradition) towards *dynamic* lexical resources (acquired from corpora, able to auto-enrich).

Such an infrastructure requires also a technical & organisational efford (crucial for a distributed & cooperative creation, management, etc. of Lexical Resources). The ELITE ("European Lexical Infrastructure and Technology") Expression of Interest submitted to the EC within the 6thFP goes in this direction.

But what to do concretely, which are the next steps and future actions? MILE can be viewed as the beginning of a common platform and one of the directions we could take is the creation of a first repository of shared lexical entries "extracted" from different available lexical resources and mapped to MILE.
Is it feasible to map different lexicon models to MILE? Is it time consuming/relatively easy?
Some possible resources to be used in the experiments are Olif, WN, Parole-Simple, ComLex, FrameNet, etc..(what else?)
Many participants volunteer to do some work in these directions, both to perfom mappings of their lexicons to MILE and to "open" core sets of lexical entries. In this ligth, it is important to understand what "*being MILE-compliant*" and "*mapping existing resources to MILE*" means in practice and to define a work methodology. It was decided that Pisa will take the initiative to start concretely this activity. The proposal made by Chu-Ren Huang was to give some coherence by choosing (if possible) lexical entries e.g. in areas related to the Olympic Games.
Starting to work in this direction is important in order to make the ELITE initiative credible.

Open, *distributed*, *shared*, *dynamic*… if these are the features of a new paradigm of LRs, we shouldn't forget the necessity to be open to a wider variety of languages.
ISLE, being an EU-US initiative, has been somehow *Euro-centric*. The next step has to be the consolidation of the already set-in-motion cooperation with Asian collegues, using a common framework and extending ISLE results in order to deal with the many particularities of the *universe* represented by Asian languages (in particular for morphology).

These topics and the others can be found in the statements given by the panelists:

"ISLE Themes and Future Initiatives" (zipped PPT format) by Nicoletta Calzolari (CPR & ILC-CNR)

"Distributed Open Lexicon Creation" (zipped PPT format) by Ralph Grishman (NYU)

"Language Aware vs. Unaware Methods" (zipped PPT format) by Stelios Piperidis (ILSP)

"Cooperation on Language Resources Sharing: Perspective from Asian Languages" (zipped PDF format) by Takenobu Tokunaga (TIT)

"Industrial Requirements" by Gudrun Magnusdottir (reference material not available)

1.3 Concluding remarks

The outcomes of the discussion stimulated by the panelists during the ending session regarding ISLE themes and future initiatives have been a concrete proposal of creating a first common repository of shared lexical entries moving, in this way, the first step towards the Open Lexical Infrastructure. In order for MILE to become the common parlance between different lexicons and models, the mapping conditions between some existing lexicons and MILE itself have to be verified. At the workshop, many people volunteered to do some work in this direction; in particular, it would be helpful if experts of a particular lexicon/dictionary can work in establishing mapping conditions between the linguistic notions of their lexicon and the MILE Basic Notions. This work can be done both at a more abstract/conceptual level and in a more formal way. To start, a grid comprising all the ISLE Basic Notions, with short descriptions, attributes and sub-elements can be prepared. The grid will comprise also a part that those who want to take npart in the "experiment" have to fill with the correspondent "notions" in their lexicons, together with a field that can be used to specify details about the mapping (there is a one-to-one mapping, a single notion in your lexicon that can be mapped to two or more distinct notions in MILE, there is only a near-mapping etc.). In the meantime, an experiment on a more formal level can be started, trying to provide an RDF description of the mapping between e.g. the ItalWordNet lexicon and MILE. The results will be circulated so that other formal mappings, hopefully, will follow.

This experiment can paves the way towards the creation of the *Open Distributed Lexical Infrastructure*, 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. 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.