The ISLE in the Ocean
Transatlantic Standards for Multilingual Lexicons (with an Eye to Machine Translation)

Nicoletta Calzolari, Alessandro Lenci, Antonio Zampolli,[1]
Nuria Bel, Marta Villegas,[2]
Gregor Thurmair[3]

[1]Istituto di Linguistica Computazionale, CNR, Pisa
Consorzio Pisa Ricerche
Università di Pisa, Dipartimento di Linguistica
Italy
{glottolo,lenci,eagles}@ilc.pi.cnr.it
[2]GILCUB (Grup Investigació Lingüística Computacional Universitat Barcelona
Spain
{nuria,tona}@gilcub.es
[3]Sail Labs, Munich
Germany
Gregor.Thurmair@sail-labs.de

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. In this paper we concentrate on the current position of the ISLE Computational Lexicon Working Group (CLWG), whose activities aim at defining a general schema for a multilingual lexical entry (MILE), as the basis for a standard framework for multilingual computational lexicons. The needs and features of existing Machine Translation systems provide the main reference points for the process of consensual definition of the MILE. The overall structure of the MILE will be illustrated with particular attention to some of the issues raised for multilingual lexicons by the need of expressing complex transfer conditions among translation equivalents

Keywords

standards, multilingual resources, computational lexicons for MT

1. Introduction

One of the crucial aspects for HLT is how to optimise the production, maintenance and extension of computational lexical resources, as well as the process leading to their integration in applications. An essential precondition to achieve these results is to establish a common and standardized framework for computational lexicon construction, which may ensure the encoding of linguistic information in such a way to grant its reusability by different applications and in different tasks. This is even more true when multilingual lexicons and machine translation (MT) are taken into consideration. Here two specific problems arise, which respectively concern *architectural* and *representational* issues: (i.) how to build new bilingual (multilingual) lexicons from available monolingual resources, and how to establish the proper relation among these two types of architectures; (ii.) how to state in the most proper way the translation correspondences among entries in the multilingual lexicon. With respect to the latter problem, 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. In turn, the representational issues are crucially connected to the architectural ones, mainly depending on how linguistic information is organized in the monolingual parts, and how it can be accessed at the multilingual layer.

The ongoing work of the Computational Lexicon Working Group (CLWG) in the ISLE project, which we illustrate in the following sections, pursues the main goal of establishing a general and consensual standardized environment for the development and integration of multilingual resources, so as to provide a satisfactory answers to the issues above. The general vision of the project 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 way the CLWG pursues this goal is 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.

This task has a crucial and special added value for MT. Although ISLE intends to address the problems of multilingual resources in its widest general aspects, MT explicitly represents the main focus of the standardization process which is being carried out by the CLWG. In fact, not only have the specific needs of MT systems been assumed as the main reference point for the CLWG work, but some of the main academic and industrial European and US actors of the MT community (Systran, Sail Labs, Lernhout & Houspie, Microsoft, LexiQuest, etc.) are also actively and directly involved in the ISLE activities.

In the next sections, we will first briefly outline the structure and general goals of the EAGLES/ISLE initiative, with special reference to the CLWG, and we will then pass to illustrate the structure and overall organization of the MILE. Finally, some specific issues raised by the process of establishing multilingual lexical correspondences will be addressed, together with an analysis of how these may affect the shape and content of the MILE.

2. The EAGLES/ISLE Initiative

The ISLE project is a continuation of the long standing EAGLES initiative (Calzolari *et al*., 1996). EAGLES stands for *Expert Advisory Group for Language Engineering Standards* and was launched within EC Directorate General XIII's Linguistic Research and Engineering programme in 1995, continued under the Language Engineering 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 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 for their long-term significance. The basic idea behind EAGLES/ISLE work is for the group to act as a catalyst in order to pool concrete results coming from current major International/National/industrial projects. Numerous theories, approaches, and systems, and relevant common practices or upcoming standards are being used as input to EAGLES/ISLE work, 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.

2.1. The Computational Lexicon Working Group

We concentrate in the following on the current position of the ISLE 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.*, 2000, 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 ISLE objective is more ambitious both in geographic scope , involving European, American and now Asian groups, and in linguistic scope, tackling the multilingual issue, which is a challenging one.

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, ii) to provide a list of cross-lingual lexical phenomena that could be used to focus the survey, and iii) to focus on MT as *de facto* the main reference application for the standardization process. Each participant engaged for surveying a number of resources. This survey is at the basis of the work of the current second phase, leading to the proposal of the MILE.

3. The structure of the MILE

The main goal of the “recommendation phase” of CLWG being the definition of a Multilingual ISLE Lexical Entry (henceforth MILE), a list of the main applications (most of which are existing MT systems) that use lexical resources was established, to focus the recommendations around them.

3.1 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, possibly targeting different application types.
- 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.
- 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 consequently ***hierarchical structure*** of the basic notions, attributes, values, etc.

3.2. The MILE overall 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 for 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, others by specifying a network of semantic relations, and others in terms of argumental frames. A way out of this theoretical variation is to augment the expressive power of an annotation scheme both *horizontally*, i.e. by distributing the annotated information over mutually independent "coding layers", and *vertically*, by further specifying the information conveyed by each such layer.

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 CLWG has also agreed that the MILE encompasses and is built on the whole monolingual entry, and will include a number of interconnected modules, which in turn further subdivide into more fine-grained structures. The three foreseen components are:

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 and subcategory
  - ➢ 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 (only for adjectives)
  - ➢ List of syntactic positions forming subcategorization frames
  - ➢ Possible syntactic realizations and grammatical functions of the positions
  - ➢ Morphosyntactic and/or lexical features (agreement, prepositions and particles introducing clausal complements)
  - ➢ Information on control (subject control, object control, etc.) 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, to characterize the actionality behaviour
  - ➢ Link to the syntactic realization of the arguments
  - ➢ Basic semantic relations between word senses:
    - o synonymy (synset)
    - o hyponymy
    - o meronymy, etc.
  - ➢ Description of word-sense in terms of more specific, various semantic/world-knowledge relations among word-senses (such as EWN relations, SIMPLE Qualia Structure, FrameNet Frame Elements, etc.)
  - ➢ Information about regular polisemous alternation in which a word-sense may enter
  - ➢ 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 (e.g. noun-noun, noun-PP, adjective noun, etc.)
- 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.

Some of these points are discussed more in detail in the following section.

4. A Multilingual Layer for the Lexicon: the ISLE approach

The line pursued by the CLWG is to define the multilingual layer of the MILE as an additional dimension on top of the monolingual ones. Related units are not modified but rather new ‘correspondence’ objects are created, pointing to already existing monolingual elements. This will to grant the maximum degree of flexibility and consistency in reusing existing monolingual resources to build new bilingual and multilingual lexicons.

Multilingual correspondences in the MILE are regarded as binary relations involving one source element and one target element. These correspondences may involve different elements, ranging from the raw surface strings, to syntactic units and semantic units, up to more abstract objects like semantic predicates, conceptual objects, etc. Correspondences can also be filtered or enriched with new information which is not present in the monolingual lexicons, but which is essential to establish multilingual correspondences.

There are several dimensions concerning the issue of correspondences, which enter into shaping their actual form:

1. *Contextuality*, i.e. the extent to which context is relevant for the description of a transfer. Two cases usually occur:

- *simple lexical transfer*, which implies replacing one lexeme of one language by one lexeme of another language.
- *complex lexical transfer, or contextual transfer*. In this case, the correspondence involves e.g. a restructuring of verbal arguments, and the multilingual module must specify how the context changes.

In order to cope with complex transfers, it is necessary to specify which are the configurational consequences of a given bilingual correspondence. This may range from a change in gender to a complete restructuring of a sentence (cf. GER. *er schwimmt gern*$_{adv}$ -> EN. *he likes*$_{vrb}$ *to swim*). As a result, the multilingual layer of the MILE will contain a whole set of conditions to express complex transformation in the SL to TL transfer, involving argument restructuring, change in the obligatoriness of positions, adjunct specifications, element addition or deletion, etc.

2. *Ambiguity*. There are two basic cases:

- in a one-to-one transfer, there is only one possibility how the target entry can be translated. This is mainly the case in special domains, especially in technical ones.
- in a one-to-many transfer, a given lexical unit can be translated in several ways. The dictionary needs to describe how the right transfer can be selected. It is an open issue how much of this selection will be in transfer and how much will be part of the analysis process. Most systems follow the approach of a simple transfer and hope that disambiguation will have taken place in analysis already. In practice, most current systems use morphosyntactic and semantic clues to identify the correct transfer relation in the one-to-many situation. As a consequence, the transfer module of an entry needs to have a *test* part to identify the correct reading of a transfer. The test part usually refers to the configuration of which the lexical unit is a part (phrase or sentence level).

3. *Lexical unit internal structure*. There are three basic cases:

- single words
- compounds (the type of German / Dutch / Finnish: agglutinated)

- multiwords (i.e. several words which together form a semantic / lexical unit); this is the most frequent case in terminology.

Not all MT systems can easily combine these cases. So, sometimes the transfer entry needs a description of what it may be the head of a multiword. Sometimes the internal structure is referred to in tests. For instance, often, functional adjectives go into compound specifiers, like GER *König* -> EN *king* but EN *royal (court)* -> GER *König* as compound specifier in *Königsho*f). Note that even systems which can handle multiwords as lexical units need expressive machinery to cover collocations, which are semantically compositional but idiosyncratic in lexeme selection).

As a consequence, the CLWG agreed to structure the multilingual module of the MILE at least three parts:
i. *test part* specifying the context which must hold for a given transfer
ii. *action part* specifying what needs to be done if this transfer is selected
iii. *typed link*s, specifying the type of the transfer link itself.
Tests and actions will be expressed by making reference to the whole representational apparatus used to characterize the monolingual linguistic information. This way, it will be possible to use all the available data structures at in order to formulate the most proper multilingual links.

5. Conclusions

In this paper we described the MILE, the multilingual lexical meta-entry proposed by the ISLE CLWG as the standard representational format for multilingual computational resources, with particular attention to the needs and requirements of MT systems. The MILE main features are i) its distributed coding architecture and ii) the multilingual layer as autonomous with respect to the monolingual modules. By doing this way, emphasis is shifted on representation modularity: 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 transfers. At the formal level, the MILE architecture will be formalized by using RDF schemata (cf. http://www.w3.org/RDF), so as to exploit the full power of this data-description language, in order to become a real common parlance for multilingual lexical resources.

Acknowledgements

We would like to thank all the members of the ISLE CLWG for their active participation to this enterprise and for their deep commitment to its success.

References

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

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

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.

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., Guimier, E., Recourcé, G., Humphreys, L., Von Rekovsky, U., Ogonowski, A., McCauley, C., Peters, W., Peters, Y., Gaizauskas, R., Villegas M. (2000) *SIMPLE Work Package 2 – Final Linguistic Specifications*, deliverable D2.2, workpackage 2, progetto LE-SIMPLE (LE4-8346).

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.