The EAGLES/ISLE initiative for setting standards: the Computational Lexicon Working Group for Multilingual Lexicons

Nicoletta Calzolari, Antonio Zampolli

Istituto di Linguistica Computazionale, CNR, Pisa, Italy
[glottolo,eagles]@ilc.pi.cnr.it

Abstract

ISLE (*International Standards for Language Engineering*), a transatlantic standards oriented initiative under the Human Language Technology (HLT) programme, is a continuation of the long standing EAGLES (*Expert Advisory Group for Language Engineering Standards)* initiative, and is carried out by European and American groups within the EU-US International Research Co-operation, supported by EC and NSF.

The objective is to support HLT R&D international and national projects, and HLT industry, by developing and promoting widely agreed and urgently demanded HLT standards and guidelines for infrastructural language resources, tools, and HLT products. ISLE targets the areas of multilingual computational lexicons (MCL), natural interaction and multimodality (NIMM), and evaluation.

We concentrate in the paper on the Computational Lexicon Working Group, describing the preliminary proposals of guidelines for the "Multilingual ISLE Lexical Entry" (MILE). We highlight some methodological principles applied in previous EAGLES, and followed in defining MILE.
We also provide a description of the EU SIMPLE semantic lexicons built on the basis of previous EAGLES recommendations. Their importance is given by the fact that these lexicons are now enlarged to real-size lexicons within National Projects in at least 8 EU countries, thus building a really large infrastructural platform of harmonised lexicons in Europe.

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 in many areas have in fact already become *de facto* widely adopted standards, and EAGLES itself is a well-known trademark and a point of reference for HLT projects and products. We stress the relevance of standardised language resources also for the humanities applications.

1 The EAGLES/ISLE Enterprise

ISLE (*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 now 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 2 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 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).

Leading industrial and academic players in the HLT field (more than 150) have actively participated in the definition of this initiative and have lent invaluable support to its execution. Moreover, the initiative is a direct result of a series of recommendations made to the EC over several years. There is a recognition that standardisation work is not only important, but is a necessary component of any strategic programme to create a coherent market, which demands sustained effort and investment.

It is important to note that the work of EAGLES (see EAGLES guidelines, http://www.ilc.pi.cnr.it/EAGLES96/home.html) must be seen in a long-term perspective. Moreover, successful standards are those which respond to commonly perceived needs or aid in overcoming common problems. In terms of offering workable, compromise solutions, they must be based on some solid platform of accepted facts and acceptable practices. EAGLES was set up to determine which aspects of our field are open to short-term *de facto* standardisation and to encourage the development of such standards for the benefit of consumers and producers of language technology, through bringing together representatives of major collaborative European R&D projects, and of HLT industry, in relevant areas. This work is being conducted with a view to providing the foundation for any future recommendations for International Standards that may be formulated under the aegis of ISO.

The current 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.

For *multilingual computational lexicons*, ISLE aims at: 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.

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; metadata descriptions for large NIMM resources; 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. 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.

In the following we concentrate on the Computational Lexicon Working Group (CLWG). We briefly present 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 is the focus of the second ongoing 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.
We describe the preliminary proposals of guidelines for the "Multilingual ISLE Lexical Entry" (MILE). We highlight some methodological principles applied in previous EAGLES, and followed in defining the MILE.
We also provide a brief description of the EU SIMPLE semantic lexicons built on the basis of previous EAGLES recommendations. We stress the relevance of standardised language resources also for the humanities applications.

2 The Computational Lexicon Working Group: an Overview

2.1 ***Standards' design and the interaction with R&D***

EAGLES work towards *de facto* standards has already allowed the field of Language Resources (LR) 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 in many areas have in fact already become *de facto* widely adopted standards, and EAGLES itself is a well-known trademark and a point of reference for HLT projects and products.

We want to highlight here the importance of having both a standard model and core language resources (LR) (e.g. corpora and lexicons) encoded according to the standard also – or even more – for applications in the humanities. It may be in fact a big advantage to have the possibility of referring to and adopting available guidelines and possibly reusing available harmonised LR (instead of starting from scratch), thus concentrating research efforts on issues more pertinent and relevant to the specific field of interest.

Existing EAGLES results in the Lexicon and Corpus areas are currently adopted by an impressive number of European - and recently also National - projects, thus becoming "the *de-facto* standard" for LR in Europe. This is a very good measure of the impact – and of the need – of such a standardisation initiative in the HLT sector. To mention just a few key examples:

- the LE PAROLE/SIMPLE resources (morphological/ syntactic/semantic lexicons and corpora for 12 EU languages, Zampolli, 1997, 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 encoding of lexicons and 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 (Monachini & Calzolari, 1996, 1999, Leech and Wilson, 1996).

The fact that the core PAROLE/SIMPLE resources are now enlarged to real-size lexicons within National Projects in at least 8 EU countries allows to have a really large infrastructural platform of harmonised lexicons in Europe, sharing the same model.

Lexical semantics has always represented a sort of *wild frontier* in the investigation of natural language, let alone when this is also aimed at implementing large scale systems based on HLT components. In fact, the number of open issues in lexical semantics both on the representational, architectural and content level might induce an actually unjustified negative attitude towards the possibility of designing standards in this difficult territory. Rather to the contrary, standardisation must be conceived as enucleating and singling out the areas in the open field of lexical semantics, that already present themselves with a clear and high degree of stability, although this is often hidden behind a number of formal differences or representational variants, that prevent the possibility of exploiting and enhancing the aspects of commonality and the already consolidated achievements.

Standards must emerge from state-of-the-art developments. With this respect, the process of standardisation, although by its own nature not intrinsically innovative, must – and actually does – proceed shoulder to shoulder with the most advanced research. Since EAGLES involves many bodies active in EU-US NLP and speech projects, close collaboration with these projects is assured and, significantly, in many cases, free manpower has been contributed by the projects, which is a sign of both the commitment of these groups/companies and of the crucial importance they place on reusability issues. Procedures have been established allowing EAGLES to access relevant material developed by EAGLES participants working in other projects. As an example, the current NSF project XMELLT on multi-words for multilingual lexicons provides valuable input to ISLE.

With no intent of imposing any constraints on investigation and experimentation, the ISLE CLWG rather aims at selecting mature areas and results in computational lexical semantics and in multilingual lexicons, which can also be regarded as stabilised achievements, thus to be used as the basis for future research. Therefore, consolidation of a standards proposal must be viewed, by necessity, as a slow process comprising, after the phase of putting forward proposals, a cyclical phase involving EAGLES external groups and projects with:

- careful evaluation and testing by the scientific community of recommendations in concrete applications;
- application, if appropriate, to a large number of European languages;
- feedback on and readjustment of the proposals until a stable platform is reached, upon which a real consensus - acquiring its meaning by real usage - is arrived at;
- dissemination and promotion of consensual recommendations.

What can be defined as *new advance* in this process is the highlighting of the areas for consensus (or of the areas in which consensus could be reached) and the gradual consciousness of the stability that evolves within the communities involved. A first benefit is the possibility, for those working in the field, of focusing their attention on as yet unsolved problems without losing time in rediscovering and reimplementing what many others have already worked on. This is true also for humanities scholars in need to use e.g. linguistically annotated corpora and computational lexicons. This is the only way our discipline can really move forward.

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) in the field of computational lexical semantics and multilingual lexicons, so that synergies will be thus enhanced, commonalities strengthened, and resources and findings usefully shared. In other terms, the process of standard definition undertaken by CLWG, and by the ISLE enterprise in general, 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.

2.2 EAGLES methodology

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. 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.

The major efforts in EAGLES concentrate on the following types of activities, which, as seen in the following, show how, on very general lines, the work is organised in the working groups.

- Detecting those areas ripe for short-term standardisation vs. areas still in need of basic research and development;
- Assessing and discovering areas where there is a consensus across existing linguistic resources, formalisms and common practices;
- Surveying and assessing available proposals or contributed specifications in order to evaluate the potential for harmonisation and convergence and for emergence of standards;
- Proposing common specifications for core sets of basic phenomena, recommendations for good practice, for standard methodologies, etc., on which a consensus can be found;
- Setting up guidelines for representation of core sets of basic features, for representation of resources, etc.;
- Feasibility studies for less mature areas;
- Suggesting actions to be taken for a stepwise procedure leading to the creation of multilingual reusable resources, elaboration of evaluation methodologies and tools, etc.

3 The ISLE Survey Phase and Recommendation Phase

3.1 ***The Survey Phase***

Following the well established EAGLES methodology, the first priority of the CLWG in the first phase of the ISLE project was to do a wide-range survey of bilingual/multilingual (or semantic monolingual) lexicons, so as to reach a fair level of coverage of existing lexical resources.

This phase is a preliminary and yet crucial step towards the main goal of the current CLWG, i.e. the definition of the "*Multilingual ISLE Lexical Entry*" (*MILE*). This is the main focus of the second phase of the project, the so called "recommendation phase", where the main objective is proposing consensual Recommendations/Guidelines. With respect to this target, one of the first objectives of the CLWG is 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 have 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.

The *Survey* of existing lexicons (see Calzolari, Grishman, Palmer, eds. 2001[1]) has been accompanied by the analysis of the requirements of a few multilingual applications, and by the parallel analysis of typical cross-lingually complex phenomena. Both these aspects have provided the general scenarios in terms of which the survey has been organised and carried out, as well as they will form the reference landmarks for the propositive phase of standard design.

The function of an entry in a multilingual lexicon is to supply enough information to allow the system to identify a distinct sense of a word or phrase in the Source Language (SL), in many different contexts, and reliably associate each context with the most appropriate translation in the Target Language (TL). The first step is to determine, of all the information that can be associated with SL lexical entries, what is the most relevant to a particular task, e.g. which notions are the more relevant to be encoded, at which descriptive level, to which elements of the entry conditions and actions for translation need to be associated, etc. The following is a (non-exhaustive) list of key applications which rely on the use of lexical resources:

- Machine Translation (MT)
- Cross-Language Information Retrieval (CLIR)
- Cross-Language Information Extraction
- Multilingual Language Generation
- Multilingual Authoring
- Speech-to-Speech Translation
- Multilingual Summarisation

We decided to focus the work of survey and subsequent recommendations around two major broad categories of application: MT and CLIR. They have partially different/complementary needs, and can be considered to represent the requirements of other application types. The multilingual applications, considered as a starting point for both phases, provide a strong applied focus in tackling multilingual lexical encoding. It is necessary in fact to ensure that any guidelines meet the requirements of industrial applications and that they are implementable.

In the preparation of the Survey, both to facilitate the identification of basic notions and the comparison of surveyed resources, and to focus on aspects of relevance to multilingual tasks, we have decided:

1. to prepare a grid for lexicon description to be used as a checklist to classify the content and structure of the surveyed resources on the basis of a number of agreed parameters of description;
2. to identify a small number of major categories of cross-lingual lexical phenomena that could be used to focus the survey, and to provide the necessary bootstrap to the propositive phase. Actually, they represent typical *hard cases*, which are helpful to highlight the various strategies that different lexicons and systems typically resort to when operating in multilingual environments. It is one of the expected by-products of the global CLWG activity to extend and refine this preliminary list, so as to provide researchers and developers with an updated map of the problematic cases in the realm of lexical information formalisation, storage, and access, together with proposals on how to tackle them.

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

In order to better analyse lexicons, we organised the Survey in three different types of resources:

- *Machine Readable Dictionaries (MRDs)*, where the rich monolingual and bilingual information is typical of the lexicographic tradition;
- *Computational Lexicons*, large lexical resources for general use where detailed morphosyntactic, syntactic and semantic information is explicit and variously represented;
- *Lexical resources for Machine Translation systems*.

Each lexicon presentation includes:

1. a description of the surveyed resource (also on the basis of the common grid, see Table 1);
2. possibly, for one or two examples from the cross-lingual lexical phenomena, an explanation of how these examples are handled by this lexicon.

The following template (drawn up starting from a preliminary list, proposed by Sue Aktins, that essentially concerned the information present in traditional dictionaries, then integrated with more detailed morphosyntactic, syntactic and semantic information, which might be available in existing computational lexicons and machine-readable dictionaries) has been used as a general grid to evaluate the content and structure of each lexical resource, verifying if the information is available and extractable and focusing on how the various types of information can be relevant to solve problems usually tackled when processing language in a bilingual or multilingual environment. The grid is obviously not intended to be complete, since it is expected that new items might be introduced as a result of the recommendation phase.

Table 1: Lexical Information in Bilingual Resources

	Entry component	
1	headword	
2	Phonetic transcription	
3	variant form	
4	inflected form	
5	Cross-reference	
6	Morphosyntactic information	
	a	Part-of-speech marker
	b	Inflectional class
	c	Derivation
	d	Gender
	e	Number
	f	Mass vs. Count
	g	Gradation
7	Subdivision counter	
8	Entry subdivision	
9	Sense indicator	
10	linguistic label	
11	Syntactic information	
	a	Subcategorization frame
	b	Obligatority of
	c	Auxiliary
	d	Light or support
	e	Periphrastic
	f	Phrasal verbs
	g	Collocator
	h	Alternations
12	Semantic information	
	a	Semantic type
	b	Argument structure
	c	Semantic relations
	d	Regular polysemy

	e	Domain
	f	Decomposition
13	Translation	
14	Gloss	
15	near-equivalent	
16	example phrase	
17	example phrase (problematic)	
18	multiword unit	
19	Subheadword *also* secondary	
20	usage note	
21	Frequency	

3.2 ***The Recommendation Phase***

The principle guiding the elicitation and proposal of MILE basic notions in the current recommendation phase will be, 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. The work of gathering descriptions and characterisations of multilingual lexical phenomena from a set of major existing lexicons, systems, dictionaries, etc., provides better ground to decide what is needed, what can be agreed on, what can be integrated in a unitary MILE, what is lacking or needs formalisation, and so on.

This method of work has proven useful in the process of reaching consensual *de facto* standards in a bottom-up approach and is at the basis also of ISLE work. There is every interest in building on existing resources, rather than starting from scratch, thus efforts must continue in this direction.

Natural language meaning has always been thought of as one of the hardest problems for standardisation. However, the increasing use of conceptual classification in the development of language technologies is rapidly changing this perception. At the same time, the growing need for dealing with semantics and contents in HLT applications is pushing towards more powerful and robust semantic components. Within the last decade, the availability of robust tools for language analysis has provided an opportunity for using semantic information to improve the performance of applications such as Machine Translation, Information Retrieval, Information Extraction and Summarisation. As this trend consolidates, the need of a protocol which helps normalise and structure the semantic information needed for the creation of reusable lexical resources within the applications of focus, and in a multilingual context, becomes more pressing. Times are thus mature to start tackling the question of how to formulate guidelines for multilingual lexical (semantic) standards.

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 are paying particular attention in the recommendation phase, and we are currently examining the extension of the EAGLES guidelines in these and other areas to propose a broad format for multilingual lexical entries which should be of general utility to the community.

In the previous EAGLES work on Lexicon Semantics (Sanfilippo et al., 1999) the following technologies were surveyed to determine which types of semantic information were most relevant:

- Machine Translation (MT)
- Information Extraction (IE)
- Information Retrieval (IR)

- Summarisation (SUM)
- Natural Language Generation (Gen)
- Word Clustering (Word Clust)
- Multiword Recognition + Extraction (MWR)
- Word Sense Disambiguation (WSD)
- Proper Noun Recognition (PNR)
- Parsing (Par)
- Coreference (Coref)

The results of the previous EAGLES survey are here summarized. Each different type of semantic information is followed by the application type in which it figures:

- BASE CONCEPTS, HYPONYMY, SYNONYMY: all applications and enabling technologies
- SEMANTIC FRAMES: MT, IR, IE, & Gen, Par, MWR, WSD, Coref
- COOCCURRENCE RELATIONS: MT, Gen, Word Clust, WSD, Par
- MERONYMY: MT, IR, IE & Gen, PNR
- ANTONYMY: Gen, Word Clust, WSD
- SUBJECT DOMAIN: MT, SUM, Gen, MWR, WSD
- ACTIONALITY: MT, IE, Gen, Par
- QUANTIFICATION: MT, Gen, Coref

It is important to notice that all of these semantic information types (except for quantification) are covered by the SIMPLE model. For this reason, the structure and the characteristics of SIMPLE (as a lexical resource designed on the basis of the EAGLES recommendations) has a crucial place in the design of the MILE. One very interesting possibility seems to be to complement WordNet-style lexicons with the SIMPLE design, thereby trying to get at a more comprehensive and coherent architecture for the development of semantic lexical resources.

MILE will also include previous EAGLES recommendations for other layers. We are evaluating the usefulness of these other layers in the multilingual perspective, e.g. for the MT and CLIR tasks. We therefore have to analyse whether existing EAGLES recommendations, or existing lexicon models, with respect to the agreed basic notions, comply with the requirements of a multilingual perspective. Differently from previous levels of description, for the multilingual task it will however most probably appear that existing models (or even the union of them) do not cover all the notions/data which are needed for multilingual tasks. In this respect, we have also to discover areas of deficiency, and highlight areas in need of further analysis. The same is true of applications: for some of the already available lexical information, current systems are not yet able to use it. Here too areas where systems could be easily improved could be spotted and put forward.

4 The SIMPLE lexicons

Given the fact that the PAROLE/SIMPLE Lexicons, based on the GENELEX model, are used and critically evaluated as a basis for the definition of the MILE, we briefly provide here some information of these resources. The design of the SIMPLE lexicons (Bel *et al.*, 2000) complies with the EAGLES Lexicon/Semantics Working Group guidelines (Sanfilippo *et al.*, 1999), and the set of recommended semantic notions.

The SIMPLE lexicons (see http://www.ub.es/gilcub/SIMPLE/simple.html for the specifications and sample lexical entries for the various languages) are built as a new layer connected to the PAROLE

syntactic layer, and encode structured "semantic types" and semantic (subcategorization) frames. They cover 12 languages (Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish). Each lexicon is based on the same common model, designed to facilitate future cross-language linking: they share the same *core ontology* and the same set of *semantic templates*.

The SIMPLE model provides the formal specification for the representation and encoding of the following information:

i) *semantic type*, corresponding to the template each Semantic Unit (*SemU*) instantiates;
ii) *domain* information;
iii) *lexicographic gloss*;
iv) *argument structure* for predicative SemUs;
v) *selectional restrictions* on the arguments;
vi) *event type*, to characterise the aspectual properties of verbal predicates;
vii) *links* of the arguments *to the syntactic subcategorization frames*, as represented in the PAROLE lexicons;
viii) *'qualia' structure*, following the Generative Lexicon (Pustejovsky, 1995), represented by a very large set of semantic relations and features;
ix) information about *regular polysemous alternation* in which a word-sense may enter;
x) information concerning *cross-part of speech relations* (e.g. *intelligent - intelligence*; *writer - to write*).
xi) *semantic relations*, such as hyponymy, synonymy, etc.

The "conceptual core" of the lexicons consists of the basic structured set of "semantic types" (the *SIMPLE ontology*) and the basic set of notions to be encoded for each sense. These notions have been captured in a common "library" of language independent *templates*, which act as "blueprints" for any given type - reflecting well-formedness conditions and providing constraints for lexical items belonging to that type.

1. TELIC [Top]
…
2. AGENTIVE [Top]
2.1. *Cause* [Agentive]
…
3. CONSTITUTIVE [Top]
3.1. *Part* [Constitutive]
3.1.1. Body_part [Part]
3.2. *Group* [Constitutive]
3.2.1. Human_group [Group]
3.3. *Amount* [Constitutive]
…
4. ENTITY [Top]
4.1. *Concrete_entity* [Entity]
4.1.1. Location [Concrete_entity]…

Figure 1. A portion of the SIMPLE Ontology.

There are three main types of formal entities:

- ***Semantic Units*** – word-senses are encoded as *Semantic Units* or *SemU*. Each SemU is assigned a *semantic type* from the Ontology, plus other sorts of information specified in the associated *template*, which contribute to the characterization of the word-sense.

- ***Semantic Type*** **-** SemUs are assigned semantic types. Each type involves structured information represented as *template*. The semantic types themselves are organized into the *Ontology* (see Figure 1), which allows for the *orthogonal organisation* of types (Pustejovsky, 1995).

- ***Template*** - a schematic structure which the lexicographer uses to encode information about a given lexical item. The template expresses the semantic type, plus other sorts of information characterising multiple dimensions of a word-sense. Templates are intended both to provide the semantics of the types (which are thus not simply labels) and to guide, harmonize, and facilitate the lexicographic work, as well as to enhance the consistency among the lexicons. A set of top common templates (about 150) have been defined during the specification phase, while the individual lexicons can add more language-specific templates as needed.

Templates provide the information that is type-defining for a given semantic type. Lexicographers can also further specify the semantic information in a SemU, by either adding other relations or features in the Qualia Structure, or by adding other types of information (e.g. domain information, collocations, etc.).

Take, for instance, the template associated with the type **`Instrument`**(on the left), followed (on the right) by the SemU for a sense of *lancet*, instantiating this template:

Usem:	1
BC number:	
Template_Type:	**[Instrument]**
Unification_path:	**[Concrete_entity | Artifact$_{\text{Agentive}}$ | Telic]**
Domain:	*General*
Semantic Class:	<Nil>
Gloss:	*//free//*
Event type:	<Nil>
Pred_Rep.:	<Nil>
Selectional Restr.:	<Nil>
Derivation:	<Nil>
Formal:	*isa* (1,<instrument>)
Agentive:	*created_by*(1,<Usem>:[Creation])
Constitutive:	*made_of*(1,<Usem>) //optional// *has_as_part*(1,<Usem>) //optional//
Telic:	*used_for*(1,<Usem>: [Event])
Synonymy:	<Nil>
Collocates:	*Collocates*(<Usem1>,…,<Usemn>)
Complex:	<Nil> //for regular polysemy//

Usem:	<lancet-1>
BC number:	
Template_Type:	**[Instrument]**
Unification_path:	**[Concrete_entity| Artifact$_{\text{Agentive}}$ | Telic]**
Domain:	*Medicine*
Semantic Class:	*Instrument*
Gloss:	a surgical knife with a pointed double-edged blade; used for punctures and small incisions
Event type:	<Nil>
Pred_Rep.:	<Nil>
Selectional Restr.:	<Nil>
Derivation:	<Nil>
Formal:	*isa* (<lancet-1>, <knife>: [**Instrument**])
Agentive:	*created_by* (<lancet-1>, <make>: [**Creation**])
Constitutive:	*made_of* (<lancet-1>, <metal>: [**Substance**]) *has_as_part* (<lancet-1>, <edge>: [**Part**])
Telic:	*used_for*(<lancet-1>, <cut>: [**Constitutive_change**]) *used_by* (<lance-1t>, <doctor>)
Synonymy:	<Nil>
Collocates:	*Collocates* (<SemU1>,…,<SemUn>)
Complex:	<Nil>

The following is a generic template for perception events (on the left), followed (on the right) by the instantiation of this template for a particular perception verb, sense 2 of *guardare* (to look):

Usem:	perception-verbs
BC number:	
Template_Type:	**[Perception]**
Unification Path	**[Psychological event]**
Domain:	*General*
Semantic Class:	*Perception*
Gloss:	//free//
Event type:	Process
Pred_Rep.:	Lex-Pred(<arg0>,<arg1>)
Selectional Restr.:	arg0=**animate**//concept// arg1:default=[**Entity**]
Derivation:	<Nil>
Formal:	*isa* (<perception-verb>,<SemU>: [**Perception**]>)
Agentive:	<Nil>
Constitutive:	*instrument* (<perception-verb>, <SemU>: [**Body-part**]) *intentionality*={yes,no}//optional//)
Telic:	<Nil>
Collocates:	*Collocates* (<SemU1>,…,<SemUn>)
Complex:	<Nil>

Usem:	<guardare-2> //look-2//
BC number:	105
Template_Type:	**[Perception]**
Unification Path	**[Psychological event]**
Domain:	*General*
Semantic Class:	*Perception*
Gloss:	Osservare con attenzione
Event type:	Process
Pred_Rep.:	guardare(<arg0>,<arg1>)
Selectional Restr.:	arg0=**animate**//concept// arg1:default=[**Entity**]
Derivation:	<Nil>
Formal:	*isa* (<guardare-2>,<percepire>: [**Perception**]>)
Agentive:	<Nil>
Constitutive:	*instrument* (<guardare-2>, <occhio>: [**Body-part**]) *intentionality* (={yes})
Telic:	<Nil>
Collocates:	*Collocates* (<SemU1>,…,<SemUn>)
Complex:	<Nil>

All the PAROLE/SIMPLE lexical information is encoded in SGML (for Italian now also in XML), and the whole PAROLE/SIMPLE model is fully represented according to a common DTD for all the 12 languages, based on the GENELEX DTD (GENELEX Consortium, 1994).

5 The structure of the prospective Multilingual ISLE Lexical Entry (MILE)

5.1 ***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 are 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 is (as said above) to discover and list the (maximal) set of (minimal/more granular) ***basic notions*** needed to describe the multilingual level, according to the ***'edited union'*** of what is available in major lexicons/models/dictionaries, at least as a starting point. Connected to this, it is expected that any MILE proposal will contain ***redundancy***. This is not problematic with regard to recommendations. It is only at the level of the lexicon instance that a lexicon builder may want to avoid redundancy, for reasons of efficiency, etc.

5.2 ***Modularity***

Modularity with respect to MILE can be thought of in 3 ways:

- 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 characterising the MILE in a certain source language. It possibly corresponds to the typology of information contained in existing major lexicons, such as PAROLE-SIMPLE, (Euro)WordNet (EWN), COMLEX, GENELEX, FrameNet, etc. With respect to these we have to i) evaluate their notions with respect to EAGLES recommendations for syntax and semantics, ii) evaluate their usefulness for multilingual tasks, iii) evaluate integrability of their notions in a unitary MILE, iv) look for deficient areas.

      Typologies of information to be part of this module possibly include (not an exhaustive list):

- Morphological layer
  - Grammatical category and subcategory
  - Gender, number, person, mood
  - 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
  - Syntactic constraints and property of the possible 'slot filler'
  - 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
  - Gloss
  - 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:
    - synonymy (synset)
    - hyponymy
    - 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)
  - Information about regular polysemous alternation in which a word-sense may enter
  - Information concerning cross-part of speech relations
  - …

As can be seen from the list above, some of these types of information try to make the senses of the MILE *explicit* through reference to formal resources such as ontologies, feature sets, lists of semantic relations, common predicates or argument structures.
A general issue to be 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.

This module will be one of the bases to define the transfer conditions, but can also be possibly detached to form a totally independent lexicon to be used in standard monolingual tasks (e.g. WSD).

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 or idiosyncratic syntactic constructions
   - Typical collocates

- Support verb construction
- Phraseological or multiwords constructions
- Compounds (e.g. noun-noun, noun-PP, adjective noun, etc.)
- Corpus-driven examples of MILE
- …

Collocations often can go beyond the sense distinctions expressed through the representation apparatus provided in (B.1.), since they convey further, more granular uses of the MILE, which simply cannot be expressed with the resources available in (B.1.). The module to deal with (B.2.) is presently missing in most of the lexicons, such as PAROLE/SIMPLE or EWN, and it is also missing in OLIF. Moreover, previous EAGLES has not carried out in-depth analyses of these issues. In this module we experiment more strongly the limits of the representation means adopted in current lexicons and models.

This module, very important multilingually, has the task both to characterise a word-sense in a more granular way and to make it possible a number of specific operations, such as WSD or translation in a specific context. Here open issues are: (i.) what is relevant, (ii.) what can be generalised and formally characterised, (iii.) what must be simply listed (but even lists may be partially categorised), (iv.) what type of representation and analysis can be provided of these phenomena (e.g. should we adopt a Mel'cuk style analysis for support verb constructions, FrameNet style description of syntactic-semantic "constructions", etc.).

The above types of information in both (B.1.) and (B.2.) may raise different issues in monolingual vs. multilingual tasks. For instance, some verb-complement pairs, although not representing particularly problematic case in the SL, may call for specific idiosyncratic transfers in the TL. Similarly, it is well-known that sense distinctions are often different in monolingual and in multilingual lexicons.

3. *Multilingual apparatus (e.g. transfer conditions and actions)* – Possible starting points for this module could be OLIF and GENELEX. The main issues are: (i.) devise the possible cases of problematic transfer (cf. for instance the list of cross-lingual linguistic phenomena of the survey phase); (ii.) identify which conditions must be expressible and which transformation actions are necessary, (iii.) select which types of information these conditions must access (within the above modules); (iv.) identify the various methods of establishing SL-->TL equivalence (e.g. translation, near equivalent, gloss, example, example + translation, etc.); (v.) examine the variability of granularity needed when translating in different languages, and the architectural implications of this.

   This module relies on both (B.1.) and on (B.2.), since it will have to access semantic and syntactic explicit information, as well as some more example-based information.

It is crucial to check if the available monolingual apparatus provided in (B.1.), augmented with (B.2.), is a good/sufficient basis, and to check specific information/data to be inserted in (or required by) the (B.3.) module. Moreover, 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.

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 1. and 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.

5.3 ***Organisational issues***

Assignments for in-depth analysis of the various information types were done, and work is now carried out by the 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).

ISLE is also implementing a simple lexicographic tool (a first prototype has just been done by N. Bel and M. Villegas), with which a sample of lexical entries will be encoded according to the MILE structure.

6 Enlargement to Asian Languages

An enlargement of the group to involve also Asian languages is going on, as an important further step. At recent CLWG Workshops at UPenn and in Pisa, representatives of Chinese, Japanese, Korean, and Thai languages were present. Also the newly formed *Asian Federation of Natural Language Processing Associations* (AFNLPA), chaired by J. Tsujii, declared interest in the ISLE standardisation initiative, and the possibility of a gradual involvement of Asian groups in the ISLE initiative is being now pursued, both through participation in future ISLE CLWG meetings/workshops and through new common initiatives.

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

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.

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

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

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

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

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

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

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

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

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

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