From Text to Content: Computational Lexicons and the Semantic Web

Alessandro Lenci[1], Nicoletta Calzolari[2], Antonio Zampolli[1,2]

[1]Università di Pisa, Dipartimento di Linguistica
via Santa Maria 36
56127, Pisa, Italy
alessandro.lenci@ilc.cnr.it

[2]Istituto di Linguistica Computazionale – CNR
Area della Ricerca di Pisa, via Moruzzi 1
56100, Pisa, Italy
{glottolo, pisa}@ilc.cnr.it

Abstract

The vision of the Semantic Web is to turn the World Wide Web into a *machine-understandable* knowledge base. According to this view, Web content is annotated with respect to particular *ontologies*, which provide the definition of the basic vocabulary and semantics of the annotations. In this paper we will argue that strengthening the synergies between ontology design and computational lexicon development is a key precondition for the Semantic Web and HLT communities to truly benefit of each other's results. In particular, we will tackle this issue by illustrating a series of requirements that computational lexicons must fulfill in order to become effective resources to contribute to implement the Semantic Web vision. These requirements will be discussed in the context of an existing infrastructure for the development of semantic language resources. Finally, we will argue that the emerging standards for the Semantic Web also provide the ground for the architecture and design of next-generation language resources.

Introduction

The vision of the Semantic Web is to turn the World Wide Web into a *machine-understandable* knowledge base, thereby allowing agents and applications to access a variety of heterogeneous resources by processing and integrating their content. Nowadays, a sheer amount of the information content available on the Web resides within natural language documents, only a part of which in English. Besides, a growing number of languages are enlarging their presence on the Web, with a relevant participation of Asian ones. Actually, this is a trend that is likely to continue, since making the Web a real global resource presupposes that users are granted the possibility to exchange information using their native tongue.

In order to make the Semantic Web a reality, it is therefore necessary to tackle the twofold challenge of *content availability* and *multilinguality* (Benjamins et al. 2002). This in turn implies fostering the way information in natural language documents is identified, extracted and explicitly represented in a such a way to become accessible by software agents. A natural convergence thus exists between the Semantic Web long-term goals and some of the core activities in the field of Human Language Technology (HLT). Multilingual semantic processing actually lies at the heart of Natural Language Processing (NLP) and Language Engineering (LE) research and technological development, since no effective text understanding can be envisaged without the proper identification and representation of the semantic content of documents encoded in different languages.

In the Semantic Web, content is annotated with respect to particular *ontologies*, which provide the definition of the basic vocabulary and semantics of the annotations. More in general, ontologies appear as key ingredients in knowledge management and content based systems, with applications ranging from document search and categorization, e-commerce, agent-to-agent communication, etc. In HLT, the task of providing the basic semantic description of words is entrusted to *computational lexicons*, which therefore represent critical information sources for most NLP systems. The availability of large-scale repositories of lexical information is in fact an essential precondition for HLT to be able to tackle the full complexity of multilingual text processing.

Ontologies also represent an important bridge between knowledge representation and computational lexical semantics, and actually form a *continuum* with semantic lexicons. In fact, they are widely used (together with lexicons) to represent the lexical content of words, and appear to have a crucial role in different natural language processing (NLP) tasks, such as content-based tagging, word sense disambiguation, multilingual transfer, etc.

Besides, one of the most widely used lexical resources, WordNet (Fellbaum 1998), is also commonly regarded and used as an ontology, as further evidence of the commonalities existing between computational lexicons and ontologies (Guarino 1998, Oltramari et al. 2002).

The main argument of this paper is that strengthening the synergies between ontology design and computational lexicon development is a key precondition for the Semantic Web and HLT communities to truly benefit of each other's results. In particular, we will tackle this issue by illustrating a series of requirements that computational lexicons must fulfill in order to become effective resources to contribute to implement the Semantic Web vision. These requirements will be discussed in the context of an existing infrastructure for the development of semantic language resources. Moreover, we will also argue for a *bi-directional* interaction between computational lexicons and the Semantic Web. Not only will computational lexicons contribute to the content-based management of information on the Web, but the emerging standards for the Semantic Web also provide the ground for the architecture and design of next-generation language resources.

A General Infrastructure for Multilingual Computational Lexicons

Computational lexicons aim at making word content machine-understandable. That is to say, they intend to provide an explicit representation of word meaning, so that it can be directly accessed and used by computational agents, such as a large-coverage parser, a module for intelligent Information Retrieval or Information Extraction, etc. In all these cases, semantic information is necessary to enhance the performance of NLP tools, and to achieve a real understanding of text content. Multilingual computational lexicons add to the representation of word meaning the information necessary to establish links among words of different natural languages, and are key components in systems for multilingual text processing, such as Machine Translation, Cross-lingual Information Retrieval, etc.

In the last decade, many activities have contributed to substantially advance knowledge and capability of how to represent, create, maintain, acquire, access, etc. large lexical repositories. These repositories are rich in linguistic knowledge, and based on best practices and standards that have been consensually agreed on or have been submitted to the international community as *de facto* standards. Core - or even large - lexical repositories have been and are being built for many languages. Besides WordNet, important examples are EuroWordNet (Vossen 1998), PAROLE (Ruimy et al. 1998), SIMPLE (Lenci et al. 2000a) in Europe, ComLex (Grishman, Macleod and Meyers 1994), FrameNet (Fillmore, Wooters and Baker 2001) in the US, among many others.

A further step and radical change of perspective is now needed in order to facilitate the integration of the linguistic information resulting from all these initiatives, to bridge the differences between various perspectives on language structure and linguistic content, to put an infrastructure into place for content description at the international level, and to make lexical resources usable within the emerging Semantic Web scenario. This objective can only be achieved when working in the direction of an integrated *open and distributed lexical infrastructure*, which is able to simultaneously tackle the following aspects:

i. to foster *the design of advanced architectures* for the representation of lexical content;
ii. to develop new methods and techniques for *the automatic acquisition of semantic knowledge* from texts and for the customization and update of lexical resources;
iii. to promote *the standardization* of various aspects of the lexicon, up to content interoperability standards.

In the sections below, we will address these points by presenting a general infrastructure for the development of semantic language resources whose main objective is to tackle the issues above in an innovative way.

Lexicon Modelling

According to a widely quoted definition (Gruber 1993), the term ontology refers to "a specification of a conceptualization", that is to say the description of "the concepts and relationships that can exist for an agent or a community of agents". An alternative and yet similar definition conceives an ontology as "a set of knowledge terms, including the vocabulary, the semantic interconnections and some simple rules of inference and logic, for some particular topic" (Hendler 2001).

Prima facie, a striking similarity exists between ontologies and computational semantic lexicons. In both cases, the goal is to carve out the shape of a particular portion of semantic space, by individuating the relevant basic elements (i.e. the concept expressed by terms or words) and the topology of relations holding among them. Actually, in computational lexical semantics ontologies are also widely used as a formal apparatus to characterize lexical content. That is to say, a set of general concepts is selected as the core repository of semantic types to classify and describe the semantic content of lexical items. This is for instance the case of the EuroWordNet Top Ontology (Rodriguez et al. 1998), which is used to describe the basic concepts of the lexical database, and the SIMPLE Core Ontology, providing the main type system to classify word senses (Lenci et al. 2000b).

Commonalities should however not overshadow the differences between ontologies and computational lexicons, nor blur the specific character of the challenge set by lexical meaning description. Differences mainly reside in the peculiar character of lexical knowledge, which computational lexicons purports at describing. Some of the main features of the latter can be described as follows:

1. lexical knowledge is inherently *heterogeneous* and *implicitly structured*. For instance, describing the semantic content of words like *part*, *material*, *link*, etc. necessarily implies to refer to their inherent relational nature. Verbs also require specific representational solutions, often quite different from the ones adopted for nouns (cf. Busa et al. 2001). In fact, the specification of the number and types of participants to the event express by the verb or the temporal properties of the event itself are crucial conditions for a satisfactory description of its meaning. Moreover, word meaning is always the product of complex dynamics: what appear in a computational lexicon must be regarded as the result of an abstraction process from the concrete and multifaceted behavior of words in texts, which in turn appear to keep on re-shaping its organization.
2. *polysemy* is a widespread and pervasive feature affecting the organization of the lexicon. The different senses of a word are only rarely separate and well-distinguished conceptual units. In a much more common situation, words have multiple meanings that are in turn deeply interwoven, and can also be simultaneously activated in the same context. *Bank* is usually quoted as a clear case of ambiguity between a location near a river, and a financial institution. The problem is that even in this case, the latter sense of *bank* is actually a constellation of different meanings: the bank-as-institution needs to be distinguished from the bank-as-a-building, and yet these two senses are clearly related in a way in which the bank of the Thames river and the Bank of England are not.
3. related to the former point, it is necessary to tackle the central issue that word senses are *multidimensional entities* that can barely be analyzed in terms of unique assignments to points in a system of concepts. As particularly argued in Pustejovsky (1995), a suitable type system for lexical representation must be provided with an unprecedented complexity of architectural design, exactly to take into account the protean nature of lexicon and its multifaceted behavior.

In the last years, the community of language resource developers is becoming aware that these issues represent essential constraints for computational lexicon design. Ignoring them or dismissing them as being purely theoretical points would end up blurring the specificity of lexical knowledge, with an negative impact on their practical usability in applications.

The need to account for the multidimensional nature of linguistic data requires the development of richer systems of semantic types, where the conceptualization expressed by word meanings must be analyzed along various orthogonal dimensions. The relational aspects of lexical items, the argument structures of predicative expressions, and the complex interplay of syntactic and semantic conditions must necessarily find a proper place within lexical architectures. Besides, the notion itself of lexical unit is not without problems, given the pervasive presence of *non-compositional aspects* in the lexicon, such as collocations, multiword expressions, idioms, etc. As a result, a suitable lexical architecture must necessarily provide a "hybrid environment", where the semantic content is represented through a careful and variously weighted combination of different types of formal entities.

An attempt to adhere to the above constraints and to meet the complexity of lexical content processing is represented by the SIMPLE model (Lenci et al. 2000b). This provides a system of semantic types for multilingual lexical encoding in which the multidimensionality of word meaning is explicitly targeted. Different aspects of the linguistic behavior of lexical items - ranging from semantic relations, to argument structure and aspect – ground the structural organization of the SIMPLE ontology. Actually, the approach specifically adopted in SIMPLE offers some relevant answers to the problems of ontology design for the lexicon, and at the same time brings to the surface other crucial issues related to the representation of lexical knowledge aiming at the development of computational lexical repositories.

The design of the SIMPLE model complies with the EAGLES Lexicon/Semantics Working Group guidelines and the set of recommended semantic notions. The model has been instantiated in semantic lexicons of about 10,000 senses covering 12 languages (Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish). Each lexicon encodes structured "semantic types" and semantic (subcategorization) frames, which are linked to the syntactic entries in the LE-PAROLE lexicon, thereby granting an optimal integration between semantic and syntactic lexical information.

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

- semantic type;
- domain information;
- lexicographic gloss;
- argument structure for predicative lexical items;

- selectional restrictions on the arguments;
- event type, to characterise the aspectual properties of verbal predicates;
- links of the arguments to the syntactic subcategorization frames, as represented in the LE-PAROLE lexicons;
- 'qualia' structure, as specified in the Generative Lexicon (Pustejovsky, 1995);
- information about regular polysemous alternation in which a word-sense may enter;
- information concerning cross-part of speech relations (e.g. *intelligent - intelligence*; *writer - to write*).
- semantic relations, such as hyponymy, synonymy, etc.

The "conceptual core" of the lexicons consists of the basic structured set of semantic types 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.

The semantic types form the *SIMPLE Core Ontology* (fig. 1) The principles of Qualia Structure have also been adopted to design the top-level ontology according to an *orthogonal organisation* of semantic types (Pustejovsky and Boguraev 1993; Pustejovsky 1995). In fact, the idea of orthogonal architectures represent an important contribution coming from the Generative Lexicon to overcome the limitations of conventional type systems, which are structured in a purely taxonomical way. Orthogonally structured ontologies essentially enrich the conventional architecture by organising the semantic types along multiple dimensions, which are given minimally by the Qualia roles.

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 sample of the SIMPLE Core Ontology.

In SIMPLE 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. We report below a schematic representation of two lexical entries (respectively for the noun *violin* and the verb *to look*), encoded according to the SIMPLE specifications:

LEMMA:	violin
SEMU_ID:	violin_1
POS:	N
GLOSS:	a type of musical instrument
DOMAIN:	music
SEMANTIC_TYPE:	instrument
FORMAL_ROLE:	isa *musical-instrument*
CONSTITUTIVE_ROLE:	has_as_part *string* made_of *wood*
TELIC_ROLE:	used_by *violinist* used_for *play*

LEMMA:	look
SEMU_ID:	look_1
POS:	V
GLOSS:	intentionally perceiving something with eyes
SEMANTIC_TYPE:	perception
EVENT_TYPE	process
FORMAL_ROLE:	isa *perceive*
CONSTITUTIVE_ROLE:	instrument *eye* intentionality = yes
PRED_REPRESENTATION:	look (Arg0: animate) (Arg1: entity)
SYN_SEM_LINKING:	Arg0 = subj_NP Arg1 = obl_PP_at

The full expressive power of the SIMPLE model is given by a wide set of features and relations, which are organized along the four Qualia dimensions proposed in the Generative Lexicon as the main axes of lexical description, i.e. *Formal Role*, *Constitutive Role*, *Agentive Role* and *Telic Role*. Features are introduced to characterize those attributes for which a closed and restricted range of values can be specified (e.g. sex={*male*, *female*}, intentionality={*yes*, *no*}, etc.). On the other hand, relations connect a given SemU to other semantic units. They are used to capture multiple aspects of word meaning, ranging from functionality (e.g. *used_for*, *used_by*), to mode of creation (e.g. *derived_from*, *created_by*) and internal constitution (e.g. *has_as_part*, *made_of*, etc.). Relations are organized along taxonomic

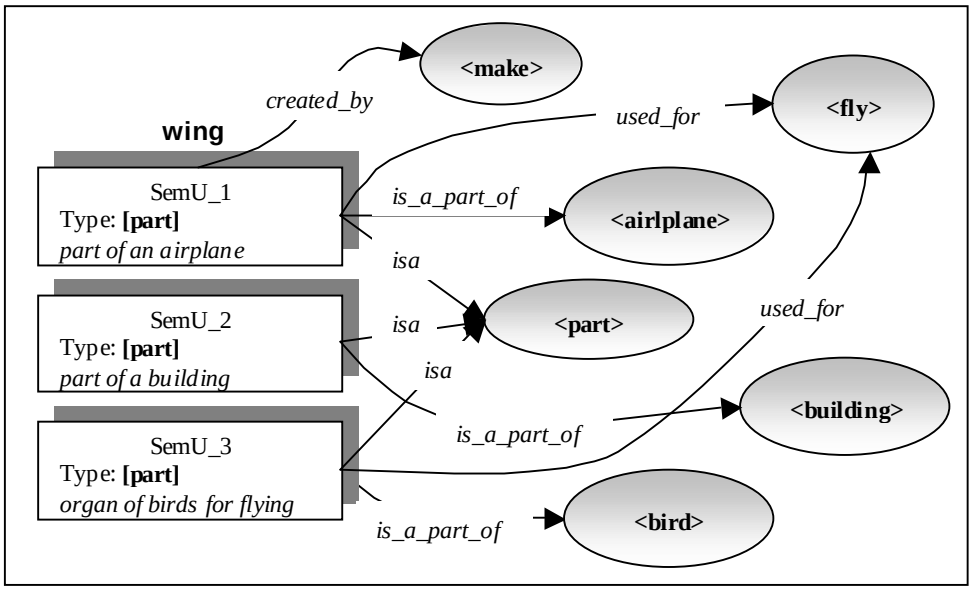

Figure 2: A semantic representation of *wing*

hierarchies, allowing for the possibility of underspecification, as well as the introduction of more refined subtypes of a given relation. Features and relations are also used to capture aspects of world knowledge relevant to characterize the semantic behavior of words. Lexical representations can thus be turned into a core inference layer licensed by words.

In SIMPLE, it is possible to capture the different semantic load of various classes of word senses, by calibrating the usage of the different pieces and types ofinformation made available by the model. For instance, figure 2 shows a possible characterization of a portion of the semantic space associated to the word *wing*. This content can be partitioned in three SemUs, that share the same semantic type (actually they are all parts), but can be nevertheless distinguished in terms of the relations they entertain with other semantic units. Thus, if on the one hand SemU_1 and SemU_3 are alike with respect to the functionality dimension (they both refer to entities used for flying), they are set apart under the constitutive aspects, since SemU_1 refers to a part of an airplane while the SemU_3 to a part of a bird., etc.

Although we are still far aside form being able to provide really satisfactory representations of word content, the SIMPLE architecture tries to approximate natural language complexity by providing a highly expressive and versatile model for language content description. The idea of a Semantic Web actually makes the understanding and modeling of the "lexical web" even more crucial, as the essential precondition for an effective natural language content processing.

Lexicon bootstrap and acquisition

The continuously changing demands for language-specific and application-dependent annotated data (*e.g.* at the syntactic or at the semantic level), indispensable for design validation and efficient software prototyping, however, are daily confronted by the *resource bottleneck*. Handcrafted resources are often too costly and time-consuming to be produced at a sustainable pace, and, in some cases, they even exceed the limits of human conscious awareness and descriptive capability. The problem is even more acutely felt for low-resource languages, since the early stages of language resource development often require gathering considerable momentum both in terms of know-how and level of funding, of the order of magnitude normally deployed by large national projects.

Secondly, computational lexicons should be rather conceived as *dynamic systems*, whose development needs to be complemented with the automatic acquisition of semantic information from texts. In fact, semantic lexical content can not be identified only through a top-down process, nor can lexical items be conceived as entities in isolation. Since conversely meanings live and arise in linguistic contexts, it is necessary to take into account how semantic information emerges from the actual textual data, and how the latter contribute to meaning formation and change. Consistently, computational lexicons are not fixed repositories of semantic descriptions, but rather provide core set of meanings that need to be customized and adapted to different domains, applications, texts, etc. This

seems to be an essential condition for language resources to be really suited to process the semantic content of documents.

Possible ways to circumvent, or at least minimize, these problems come from the literature on *automatic knowledge acquisition* and, more generally, from the machine-learning community. Of late, a number of machine learning algorithms have proved to fare reasonably well in the task of incrementally bootstrapping newly annotated data from a comparatively small sample of already annotated resources. Another promising route consists in automatically tracking down recurrent knowledge patterns in relatively unstructured or implicit information sources (such as free texts or machine readable dictionaries) for this information to be molded into explicit representation structures (e.g. subcategorization frames, syntactic-semantic templates, ontology hierarchies etc.). In a similar vein, several strategies have been investigated aimed at merging or integrating structured information sources into a unitary comprehensive resource, or at customizing general-purpose knowledge-bases for them to be of use in more technical domains. Recent contributions to this topic can be found in Lenci, Montemagni and Pirrelli (2001) and (2002), where various techniques of lexical information automatic acquisition are represented.

Actually, the need for turning symbolic explicit representations of semantic knowledge into more dynamic structures is also receiving increasing attention in the field of *ontology learning* (Staab et al. 2000). The power (and limit) of ontologies in fact lies in their ability to provide a *snapshot* of a given domain of knowledge. The basic challenge is then turning this snapshot into a dynamic and evolving structure, which might really be able to tackle the complex processing of word meaning acquisition, change and extension. Issues such as the automatic enrichment and customisation of ontologies are high on knowledge engineers' agendas, and interesting interactions with the study of the cognitive dynamics of concept formation and change can be envisaged.

Another interesting line of research is given by the contextual approaches to word meaning for NLP applications (Pereira Tishby and Lee 1993, Lin 1998, Allegrini, Montemagni and Pirrelli 2000). The central claim here is that "substitutability without loss of plausibility is an important factor underlying judgements of semantic similarity" (Miller and Charles, 1991). This interest has both practical and theoretical reasons. For our present concerns, suffice it to point out that the approach has the potential of shedding light on issues of context-sensitive semantic similarity, while getting around the bottle-neck problem of sparse data. From this perspective, meaning similarity is not dependent on the way concepts are defined in the first place (as lists of both necessary and sufficient defining properties), but acts as a determinant of concept formation, by "discretizing" the contextual space defined by the vectorial representations of word uses.

All these attempts at bootstrapping lexical knowledge are not only of practical interest, but also point to a bunch of germane theoretical issues. Gaining insights into the deep interrelation between representation and acquisition issues is likely to have significant repercussions on the way linguistic resources will be designed, developed and used for applications in the years to come. As the two aspects of knowledge representation and acquisition are profoundly interrelated, progress on both fronts can only be achieved, in our view of things, through a full appreciation of this deep interdependency.

International Standards for Lexical Resources

Optimizing the production, maintenance and extension of computational lexical resources, as well as the process leading to their integration in applications is of the utmost importance. A crucial 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. Thus, enhancing the sharing and reusability of multilingual lexical resources can be reached by promoting the definition of a common parlance for the community of computational lexicon developers and users. This is parallel to the growing efforts to foster ontology sharing and standardization, which are acknowledged as essential steps on the way towards the Semantic Web.

The SIMPLE model we presented above is directly related to the standardization initiative promoted by the ISLE Computational Lexicon Working Group (CLWG). The ISLE[1] (*International Standards for Language Engineering*) project is a continuation of the long standing EAGLES initiative (Calzolari, McNaught and Zampolli 1996).[2] ISLE 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.

EAGLES work towards *de facto* standards has already allowed the field of Language Resources (LR) to establish broad consensus on critical issues for some well-established areas, providing thus a key opportunity for

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

[2] 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 1993, continued under the Language Engineering programme, and now under the Human Language Technology (HLT) programme as ISLE, since January 2000.

further consolidation and a basis for technological advance. EAGLES previous results have already become *de facto* standards. Existing EAGLES results in the Lexicon and Corpus areas are currently adopted by a number of European - and recently also National – projects (e.g. LE-PAROLE and LE SIMPLE), thus becoming "the *de-facto* standard" for LR in Europe.

The ISLE Computational Lexicon Working Group is committed to the consensual definition of a standardized infrastructure to develop multilingual resources for HLT applications, with particular attention to the needs of Machine Translation (MT) and Crosslingual Information Retrieval (CLIR) systems. Compared with other standardization initiatives active in this field (e.g. OLIF-2; cf. Lieske, McCormick and Thurmair 2001), the original character of ISLE resides in its specifically focusing on the *grey area* of HLT where well-assessed language technology meets more advanced levels and forms of linguistic description. In particular, various aspects of lexical semantics, although still part of ongoing research, are nevertheless regarded by industrials and developers as the "next-step" in new generation multilingual applications. Standard definition in this area thus means to lay a first bridge between research in multilingual resource development and its exploitation in advanced technological systems.

Lexical semantics has always represented a "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 - in the open field of lexical semantics - the areas 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 ISLE CLWG adheres to the leading methodological principle that the process of standardization, although by its own nature not intrinsically innovative, *must – and actually does – proceed shoulder to shoulder with the most advanced research*. Consistently, the ISLE standardization process pursues a twofold objective:

1. defining standards both at the content and at the representational level for those aspects of computational lexicons which are already widely used by applications;
2. proposing recommendations for the areas of computational lexical semantics which are still in the "front line" of ongoing research, but also appear to be ready for their applicative exploitation, and are most required by HLT systems to achieve new technological leap forwards.

This double perspective is one of the peculiar features of the ISLE activities, and contributes to its added value with respect to other current standardization initiatives. This way, ISLE intends on the one hand to answer to the need of fostering the reuse and interchange of existing lexical resources, and on the other hand to enhance the technological transfer from advanced research to applications.

The 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 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 proposals.

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 re-implementing what many others have already worked on. Useful indications of *best practice* will therefore come to researchers as well as resource developers. This is the only way our discipline can really move forward.

Finally, one of the targets of standardization, and actually one of the main aims of the ISLE 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. The ISLE-CLWG pursues this goal by designing MILE (*Multilingual ISLE Lexical Entry*), a general schema for the encoding of multilingual lexical information. This has to be intended as

a meta-entry, acting as a common representational layer for multilingual lexical resources.

In its general design, MILE is envisaged as a highly *modular* and *layered* architecture, as described in Calzolari et al. (2001). Modularity concerns the "horizontal" MILE organization, in which independent and yet linked modules target different dimensions of lexical entries. On the other hand, at the "vertical" level, a layered organization is necessary to allow for different degrees of granularity of lexical descriptions, so that both "shallow" and "deep" representations of lexical items can be captured. This feature is particularly crucial in order to stay open to the different styles and approaches to the lexicon adopted by existing multilingual systems.

At the top level, MILE includes two main modules, *mono-MILE*, providing monolingual lexical representations, and *multi-MILE*, where multilingual correspondences are defined. With this design choice the ISLE-CLWG intends also to address the particularly complex and yet crucial issue of multilingual resource development through the integration of monolingual computational lexicons. Mono-MILE is organized into independent modules, respectively providing *morphological*, *syntactic* and *semantic* descriptions. The latter surely represents the core and the most challenging part of the ISLE-CLWG activities, together with the two other crucial topics of *collocations* and *multi-word expressions*, which have often remained outside standardization initiatives, and nevertheless have a crucial role at the multilingual level. This bias is motivated by the necessity of providing an answer to the most urgent needs and desiderata of next generation HLT, as also expressed by the industrial partners participating to the project. With respect to the issue of the representation of multi-word expressions in computational lexicons, the ISLE-CLWG is actively cooperating with the NSF sponsored XMELLT project (Calzolari et al. 2002).[3]

Multi-MILE specifies a formal environment for the characterization of multilingual correspondences between lexical items. In particular, source and target lexical entries can be linked by exploiting (possibly combined) aspects of their monolingual descriptions. Moreover, in multi-MILE both syntactic and semantic lexical representations can also be enriched, so as to achieve the granularity of lexical description required to establish proper multilingual correspondences, and which is possibly lacking in the original monolingual lexicons.

According to the ISLE approach, monolingual lexicons can thus be regarded as *pivot lexical repositories*, on top of which various language-to-language multilingual modules can be defined, where lexical correspondences are established by partly exploiting and partly enriching the monolingual descriptions. This architecture guarantees the independence of monolingual descriptions while allowing for the maximum degree of flexibility and consistency in reusing existing monolingual resources to build new bilingual lexicons.

The *MILE Data Model* is intended to provide the common representational environment needed to implement such an approach to multilingual resource development, with the goal of maximizing the reuse, integration and extension of existing monolingual computational lexicons. The main objective is to provide computational lexicon developers with a formal framework to encode MILE-conformant lexical entries. Some of the main features of the MILE Data Model are reported below:

- it is based on the experience derived from existing computational lexicons (e.g. LE-PAROLE, SIMPLE, WordNet, EuroWordNet, etc.);
- it is structured according to the entity-relationship schema;
- it is geared towards the development of large-scale lexical databases;
- it is open to various types of users and geared towards customization;
- it is based on a *distributed* architecture
- it is open towards the use of RDF descriptions to characterize lexical objects.

The MILE Data Model will include the following four main components, as illustrated in figure 3:

1. an XML DTD formalizing the *MILE Entry Skeleton* according to an entity-relationship schema;
2. a definition of *MILE Lexical Data Categories*, forming the basic components of MILE conformant entries;
3. a first repository of *MILE Shared Lexical Objects*, instantiating the MILE Lexical Data Categories, to be used to build in an easy and straightforward way lexical entries.
4. the *ISLE Lexicographic Station*, which will map the MILE entity-relationship model into a relational database, and will also include a GUI to input, browse and query the data in a user-friendly way.

The MILE Lexical Data Categories will define the lexical objects to be used in building MILE conformant lexical entries, according to the MILE Entry Skeleton. Lexical objects include semantic and syntactic features, semantic relations, syntactic constructions, predicate and arguments, etc. The specifications of the Lexical Data Categories will act as class definitions in an object-oriented language. Lexical Data Categories will be organized in a hierarchy and will be defined using RDF schema (Brickley and Guha 2000), to formalize their

[3] "Cross-lingual Multiword Expression Lexicons for Language Technology", Nancy Ide, Vassar, PI; NSF Award No. 9982069, May 1,2000 – December 31, 2001.

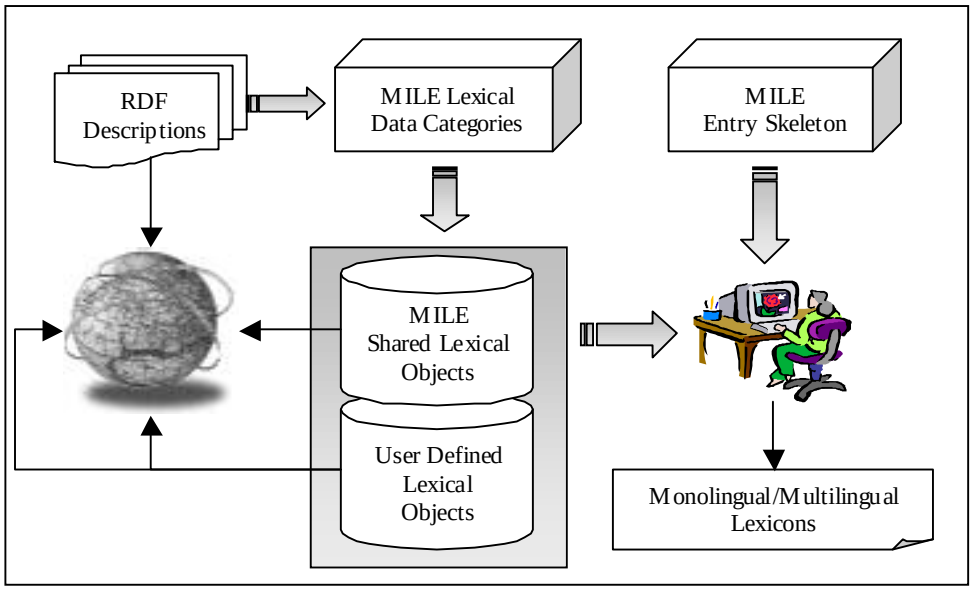

Figure 3: The MILE Data Model

properties and make their "semantics" explicit.

The MILE Shared Lexical Objects will represent instances of MILE Lexical Data Categories. The will form a first repository of recommended lexical objects, selected for their lexicographic relevance or because they represent *de facto* standards in the NLP community. This repository might include the SIMPLE Core Ontology or some of the semantic relations adopted in EuroWordNet. Users will be able to define new instances of lexical objects for their lexicon or language specific needs. This way, both an the monolingual and at the multilingual level (but with particular emphasis on the latter), ISLE intends to start up the incremental definition of a more object oriented architecture for lexicon design. Developers will be able to develop their own lexicon project either by selecting some of the MILE Shared Lexical Objects or by defining new MILE conformant objects, which in turn might then enrich the common core if they reach a certain amount of consensus in the field. Lexical objects will be identified by a URI and will act as common resources for lexical representation, to be in turn described by RDF metadata. This way ISLE intends to foster the vision open and distributed lexicons, with elements possibly residing in different sites on the Web. RDF descriptions and common definition will grant lexical content interoperability, enhancing the re-use and sharing of lexical resources and components.

Conclusions

Semantic content processing lies at the heart of the Semantic Web enterprise, and requires to squarely address the complexity of natural language. Existing experience in language resource development proves that such a challenge can be tackled only by pursuing a truly interdisciplinary approach, and by establishing a highly advanced environment for the representation and acquisition of lexical information, open to the reuse and interchange of lexical data.

Coming from the experience gathered in developing advanced lexicon models such as the SIMPLE one, and along the lines pursued by the ISLE standardization process, a new generation of lexical resources can be envisaged. These will crucially provide the semantic information to necessary allow for effective content processing. On the other hand, they will in turn benefit from the Semantic Web itself. Thus, it is possible to state the existence of a *bi-directiona*l relation between the Semantic Web enterprise and computational lexicon design and construction. In fact, the Semantic Web is going to crucially determine the shape of the language resources of the future. Semantic Web emerging standards, such as ontologies, RDF, etc., allow for a new approach to language resource development and maintenance, which is consistent with the vision of an open space of sharable knowledge available on the Web for processing

References

Allegrini, P.; Montemagni, S.; and Pirrelli, V. 2000. Learning Word Clusters from Data Types. In Proceedings of Coling 2000, Saarbruecken, Germany, July 2000.

Benjamins, V. R.; Contreras, J.; Corcho, O.; and Gómez-Pérez, A. 2002. Six Challenges for the Semantic Web. In Proceedings of SemWeb@KR2002 Workshop, 19-20 April 2002, Toulose.

Brickley, D. and Guha, R.V. 2000. Resource Description Framework (RDF) Schema Specification. W3C Proposed Recommendation; http://www.w3.org/TR/rdfschema/.

Busa, F.; Calzolari, N.; Lenci, A.; and Pustejovsky J. 2001. Building a Semantic Lexicon: Structuring and Generating Concepts, in Bunt H.; Muskens R.; and Thijsse E. eds. *Computing Meaning Vol. II*, Dordrecht, Kluwer: 29-51.

Calzolari, N.; McNaught, J.; and Zampolli, A. 1996. *EAGLES Final Report: EAGLES Editors' Introduction*. EAG-EB-EI, Pisa, Italy.

Calzolari, N.; Lenci, A.; Zampolli, A.; Bel, N.; Villegas, V.; Thurmair G. 2001. The ISLE in the Ocean. Transatlantic Standards for Multilingual Lexicons (with an Eye to Machine Translation). In Proceedings of MT Summit VIII, Santiago De Compostela, Spain.

Calzolari, N.; Fillmore, C. J.; Grishman R.; Ide, N.; Lenci A.; MacLeod, C.; and Zampolli A. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. In Proceedings of LREC 2002, Las Palmas, Spain.

Fellbaum, C. (ed.) 1998. *WordNet. An Electronic Lexical Database*, Cambridge, Cambridge, The MIT Press.

Fillmore, C. J.; Wooters, C.; and Baker, C. F. 2001. Building a Large Lexical Databank Which Provides Deep Semantics. In Proceedings of the Pacific Asian Conference on Language, Information and Computation, Hong Kong.

Grishman, R.; Macleod C.; and Meyers A. 1994. COMLEX Syntax: Building a Computational Lexicon. In Proceedings of Coling 1994, Kyoto.

Gruber, T. R. 1993. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220.

Guarino, N. 1998. Some Ontological Principles for Designing Upper Level Lexical Resources. In Proceedings of LREC1998, Granada, Spain: 527-534.

Hendler, J. 2001. Agents and the Semantic Web. *IEEE Intelligent Systems Journal*, March/April.

Lenci, A.; Bel, N.; Busa, F.; Calzolari, N.; Gola, E.; Monachini, M.; Ogonowsky, A.; Peters, I.; Peters, W.; Ruimy, N.; Villegas, M.; and Zampolli, A. 2000a. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13 (4): 249-263.

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.; and Villegas M. 2000b. *SIMPLE Work Package 2 – Final Linguistic Specifications*, deliverable D2.2, workpackage 2, LE-SIMPLE (LE4-8346).

Lenci, A.; Montemagni, S.; and Pirrelli, V. eds. 2001. *Proceedings of the Workshop on Semantic Knowledge Acquisition and Categorisation*, XIII ESSLLI, Helsinki, Finland, 13-24 August.

Lenci, A.; Montemagni, S.; and Pirrelli, V. eds. 2002. *Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, LREC2002, Las Palmas, Spain, 1 June.

Lieske, C.; McCormick, S.; and Thurmair, G. 2001. The Open Lexicon Interchange Format (OLIF) Comes of Age. In Proceedings of the MT Summit VIII, Santiago de Compostela, Spain.

Lin D. 1998. Automatic Retrieval and Clustering of Similar Words. In Proceedings of COLINGACL'98, Montreal, Canada, August 1998.

Miller, G.A. and Charles, W.G. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6 (1): 1-28.

Oltramari, A.; Gangemi, A.; Guarino, N.; and Masolo, C. 2002. Restructuring WordNet's Top-Level: The OntoClean Approach. In Proceedings of LREC2002 (OntoLex workshop), Las Palmas, Spain.

Pereira, F.; Tishby, N.; and Lee, L. 1993. Distributional Clustering Of English Words. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics: 183-190.

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

Pustejovsky, J. and Boguraev, B. 1993. Lexical Knowledge Representation and Natural Language Processing. *Artificial Intelligence*, 63: 193-223.

Rodriguez, H.; Climent, S.; Vossen, P.; Bloksma, L.; Peters, W.; Alonge, A.; Bertagna, F.; and Roventini, A. 1998. The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. *Computers and the Humanities*, 32: 117-152.

Ruimy, N.; Corazzari, O.; Gola; E., Spanu; A., Calzolari; N.; and Zampolli, A. 1998. The European LE-PAROLE Project: The Italian Syntactic Lexicon. In Proceedings of the LREC1998, Granada, Spain: 241-248.

Staab, S.; Maedche, A.; Nedellec, C.; and Wiemer-Hastings P. eds. 2000. *Proceedings of the First ECAI Workshop on Ontology Learning OL'2000*, Berlin, Germany, 25 August.

Vossen, P. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, 32: 73-89.