A New Paradigm of an Open Distributed Language Resource Infrastructure: the Case of Computational Lexicons

Nicoletta Calzolari and Claudia Soria

Istituto di Linguistica Computazionale del CNR
Via Moruzzi 1
56010 Pisa, Italy
{nicoletta.calzolari,claudia.soria}@ilc.cnr.it

Abstract

The need of ever growing language resources calls for us to propose and promote a change in the overall model of how to build, maintain and share Language Resources (LRs). A new paradigm is required, under the form of an open, distributed and collaborative language infrastructure, based on open content interoperability standards. Existing experience in LR development proves that such a challenge can be tackled only by pursuing - on the organisational side - a truly interdisciplinary and cooperative approach, and by establishing - on the technical side - a highly advanced environment for the representation and acquisition of linguistic information, open to the reuse and interchange of linguistic data. In this paper we explore the idea of this new paradigm for language resources applied to the case of computational lexicons.

Introduction

Computational lexicons aim at providing an explicit representation of word meaning, so that it can be directly accessed and used by computational agents. In the last decade, many activities at European level and world-wide have contributed to substantially advance knowledge and capability of how to represent, create, maintain, acquire, access, and share large lexical repositories. These repositories are rich in linguistic knowledge (and often in world knowledge), and many are 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 European (and non EU) languages. Besides WordNet, important examples are EuroWordNet (Vossen, 1998), PAROLE (Ruimy et al., 1998), SIMPLE (Lenci et al. 2000) in Europe, ComLex (Grishman, Macleod, and Meyers, 1994), FrameNet (Fillmore, Wooters, and Baker, 2001) in the US, among the others. However, most existing lexical resources do not have enough coverage, not only for practical reasons, but for more structural and inherent reasons. No "static" resource can ever be adequate and satisfying, neither in extension (since it cannot cover new formations, or all the possible domains) nor in depth (since it cannot provide all the necessary and useful linguistic information, not even for the existing lexical entries). It is thus essential that lexicons are built and maintained in such a way to integrate and harmonize the linguistic information coming form different sources, bridging across the differences in perspectives on language structure and linguistic content, merging existing information sources while preserving their integrity and specificity. A further step and radical change of perspective is now needed in order to facilitate this integration and to put an infrastructure into place for content description and content interoperability at European level and beyond, as well as 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, based on open content interoperability standards, where not only the linguistic experts can participate, but which includes designers, developers and users of content encoding practices, and also many members of the society.

A New Paradigm for Computational Lexicons Development and Sharing

A new paradigm of an Open Lexical Infrastructure (OLI) requires radically new approaches at various levels. In this new paradigm of lexicon development, interoperability becomes an essential keyword, not solely at the level of technical encoding, formats, and linguistic encoding. Interoperability is also needed among different communities sharing similar needs, and, maybe most critically, at the organizational, architectural and engineering levels. In the next sections we explore some of the most critical issues.
Architectural, Representational and Content Issues

Standardization and Lexical Modelling. For the resources to be multilingual and shareable, the harmonization task between the different resources becomes a very important and extensive one. The standardization effort must involve the extension and integration of existing and emerging open lexical and terminological standards and best practices such as EAGLES (Sanfilippo et al. 1999), ISLE (Calzolari, Lenci, and Zampolli 2001), TEI, OLIF, Martif (ISO 12200), Data Categories (ISO 12620), ISO/TC37/SC4. Initiatives towards the creation of lexical metadata such as IMDI, Dublin Core and OLAC must be taken into account. It is already clear that there is a great convergence among these initiatives, and synergies can be "easily" created. For example, the Computational Lexicon Working Group of ISLE has defined a general schema for the encoding of multilingual lexical information, the MILE (Multilingual ISLE Lexical Entry, Calzolari et al. 2003). This has to be intended as a meta-entry, acting as a common representational layer for multilingual lexical resources. Main related concepts are the following: The MILE Lexical Data Categories provide the lexical objects (such as syntactic and semantic features, semantic relations, syntactic constructions, predicates and arguments etc.) that are the basic components of MILE-conformant lexical entries. Lexical Data Categories are organized in a hierarchy and defined using RDF schema to formalize their properties and make their "semantics" explicit (Ide, Lenci, and Calzolari 2003). The MILE Shared Lexical Objects instantiate the MILE Lexical Data Categories, to be used to build in an easy and straightforward way lexical entries. These include main syntactic constructions, basic operations and conditions to establish multilingual links, macro-semantic objects, such as lexical conceptual templates acting as general constraints for the encoding of semantic units.

The MILE Entry Skeleton, formalized as an XML DTD, is an Entity Relationship model that defines the general constraints for the construction of "ideal" multilingual entries, as well as the grammar to build the whole array of lexical elements needed for a given lexical description. A core set of consensual basic lexical notions can be implemented as basic shared lexical "objects". Lexical data categories and objects will be identified by an URI and will act as a common resource for lexical representation, to be in turn described by RDF metadata. They can be seen as a step in the direction of simplifying and improving the usability of the MILE recommendations, and of opening lexicon building to a cooperative effort. Lexicon (or applications) developers will be encouraged to use these lexical objects to build and target lexical data at a high level of abstraction, i.e. for the description, creation, management, and delivery of their resources, therefore creating semantic interoperability. A test on the potentialities of MILE as a standard environment for computational lexicons is described in (Bertagna et al. 2004).

Lexicon Integration and Interoperability. The model of open data categories will foster language resources integration and interoperability, through links to these common standards. New objects can be progressively created and linked to the core set.

We foresee an increasing number of well-defined linguistic data categories and lexical objects stored in open and standardised repositories which will be used by users to define their own structures within an open lexical framework. It is this re-usage of shared linguistic objects which will link new contents to the already existing lexical objects, while enabling shareability of distributed lexicon portions. The design of an abstract model of lexicon architecture will ensure a flexible model while working with a core set of lexical data categories. It will guarantee freedom for the user to add or change objects if that is deemed necessary, will provide an evaluation protocol for the core standard lexical data categories, and will require verification methods for the integration of new objects. We are currently working towards this end in two closely related frameworks, namely the Working Group 4 on Lexical Standards of ISO/TC 37/SC 4 and the recently approved EU project LIRICS.

From Static to Dynamic Lexicons. The promotion of a change of perspective on lexicons as static resources towards dynamic entities, whose content is co-determined by automatically acquired linguistic information from text corpora and from the web, is deemed essential. We stress the need of using (semi-)automatic or machine aided methods wherever possible in resource work. Semantic Web initiatives are also focusing on the building of ontological representations from texts, and in this respect show a large amount of conceptual overlap with the notion of a dynamic lexicon. In automating the information extraction process, lexical resources are a fundamental piece, and acquisition may require cycles of adjustments in the lexical modelling. The acquisition tools must be able to increase the repository with new words/terms, possibly their definitions, domain, etc., from digital material, to learn concepts from text – including automatic multilingual thesaurus building, and to tailor resources to specific needs. This implies, for the machine learning community, developing new and stronger algorithmic methodologies to model textual statistics, and integrating them with traditional NLP tools. In a web-service oriented implementation of this idea, agents will look for examples, identify uses in monolingual/multilingual web texts for glossary creation. This will ensure also virtual links between lexicons and examples (corpus/web samples, image samples, clips and videos, etc.), pushing towards new and innovative types of lexicons, i.e. a sort of "example based living lexicons", linguistic objects that participate of properties of both lexicons and corpora.
A Cooperative Model. This vision will pave the way to the realisation of a common platform for interoperability between different fields of linguistic activity - such as lexicology, lexicography, terminology - and Semantic Web development. The platform will provide a flexible common environment not only for linguists, terminologists and ontologists, but also for content providers and content management software vendors. The lexicons may be distributed, i.e. different building blocks may reside at different locations on the web and be linked by URLs. This is strictly related to the Semantic Web standards (with RDF metadata to describe lexicon data categories), and will enable users to share lexicons and collaborate on parts of them. Overall, the lexicons will perform the bridging function between documents and conceptual categorisation. The descriptive framework and the tools must be tailored to the needs of the creation of both general vocabularies and terminological repositories for different worlds and domains. The common conceptual model within the envisaged architecture will ensure content interoperability between texts, multimodal documents, lexicons and ontologies.

Organizational Issues

Integration of Communities. The effort of making millions of “words” available for dozens of languages is something that no individual or small group is able to afford. It is already proved by a number of EU projects that lexicon building and maintenance can be achieved in a cooperative way. We claim that the field of computational lexicons is mature enough to broaden and open the concept of cooperative effort to a much larger set of communities, so that lexicon creation, updating and maintenance will be no longer only a lexicographers' task, but will involve broad groups of experts, and - why not? – also “non-experts” and the general public. Lexicons undoubtedly form an essential component and a building block of great impact to make the vision of a pervasive Information Infrastructure and of the Semantic Web a reality. The availability of large computational lexicons is an essential prerequisite to the twofold challenge of digital content and multilinguality. Semantic Web developers will need repositories of words and terms, as well as knowledge about their relations within language use and ontological classification. However, the cost of adding structured and machine-understandable lexical information can be again a bottleneck, i.e. one of the factors that delay semantic web full deployment. Unfortunately, linguists alone will not be able to solve this. To achieve a real break-through, a radical shift in the lexical paradigm - whereby many participants add linguistic content descriptions in an open distributed lexical framework – is required to make the Web usable.

International Cooperation and Organization. International cooperation will become certainly the most important factor for the overall field of language resources – and consequently of computational lexicons - in the next years. A report produced by ELDA presents an analysis of several organisational frameworks, focussing on funding and organisational procedures for providing LRs (Calzolari et al. 2004). The pre-requisites to be addressed for the production of interoperable LRs in a cooperative framework belong to different layers: technical (specifications), validation (quality assessment), legal, commercial. In order to fill the gaps in terms of LRs, cooperation on all combined - organisational, funding, technical and commercial - issues appears to be necessary. To strengthen such a cooperation, there is no doubt that a supra-national effort in coordinating this cooperation is required.

Validation and Quality Control Activities. Quality control and validation become very important issues to ensure the quality of cooperatively built resources. There is obviously always a greater danger of unreliability in semi-automatic harvesting and dynamic extension of the lexicon. The opportunity that everyone could access existing data to put in new information, requires careful consideration on how quality can be ensured. Verification methods are required for the integration of new lexical data categories, and validation protocols need to be designed and adopted (e.g. a mechanism of certificates of validity can be designed, according to different criteria).

Market-related Issues, Access and Pricing. Access and pricing policies must be carefully considered and designed, and may be different for different uses and users types. Access policy requires also study of national versions of copyright law. A business model has to be studied, based on the idea of web-based services.

Conclusions

We claim that it is feasible and timely - and indeed critically needed - to set up a very large world-wide initiative, within which a large group of lexicon experts/providers is a core component. This initiative must build prominently on existing or future (possibly national) lexical initiatives and integrate the human R&D resources along this dimension. However, an essential aspect to ensure an integrated basis is to enhance the interchange and cooperation among many communities that act now separately, such as Language Resources developers, Terminology, Semantic Web and Ontology experts, content providers. This is the challenge for the next years, for a usable and useful lexical scenario in the global network.

References


