A Computational Multi-Layered Italian Lexicon for HLT Applications

Nilda Ruimy
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1 – 56124 Pisa, Italy
nilda.ruimy@ilc.cnr.it

Abstract
In this paper, the largest computational, lexical knowledge base of Italian language is presented. PAROLE-SIMPLE-CLIPS is a general-purpose lexicon in which 55,000 words are encoded at four different levels of linguistic description. It is based on the international standards set out in the PAROLE-SIMPLE model that was consensually agreed on and adopted for creating lexicons of 12 European languages. The original and innovative features of the lexical model are outlined. The richness of the information types encoded at the four descriptive levels is illustrated with a special focus on those that are particularly specific to the underlying model, viz. the GL inspired representation of semantic knowledge, with the expressive power of the *Extended Qualia Structure*, the encoding of semantic argument structure and the connection between syntactic and semantic information. Finally, the way such a wealth of highly structured, granular and innovative data can be exploited in HLT applications is mentioned.

1 Introduction

The importance of accessing large repositories of lexical knowledge for HLT applications is now widely acknowledged. To come up to the expectation of the scientific community, the design of a lexical model and the development of large-scale, multi-layered, general-purpose and harmonized lexicons for all European languages were performed in the framework of the PAROLE and SIMPLE EC projects. For the Italian language, the lexicon development effort was then carried on in the context of the national project *Corpora e Lessici dell'Italiano Parlato e Scritto* (CLIPS).[1]

The resulting PAROLE-SIMPLE-CLIPS Italian lexicon, developed between 1996 and 2003, was therefore entirely built in accordance with the international standards set out in the PAROLE-SIMPLE lexical model (Ruimy *et al.* 1998; Lenci *et al.* 2000) that was consensually adopted for the representation of lexical information in 12 European languages of very different types.

[1] The CLIPS project (2000-2003), funded by the Ministero dell'Istruzione, dell'Università e della Ricerca, was launched by Antonio Zampolli from the Institute for Computational Linguistics, also on behalf of the Consorzio Pisa Ricerche.

In this paper, the theoretical and representational framework of the lexical resource is outlined. An overview of the lexicon general architecture is provided. The syntactic and semantic levels and their information content are illustrated. Focus is put on some particularly relevant aspects of the semantic representation.

2 The Lexical Model: Conceptual and Representational Framework

The PAROLE-SIMPLE conceptual model is grounded on the EAGLES[2] recommendations that constitute de facto standard for computational lexicons in Europe, on the extended GENELEX model as well as on the results of EuroWordNet, ACQUILEX and DELIS European projects. The theoretical approach to the representation of semantic information is essentially grounded on a revised version of some fundamental aspects of the Generative Lexicon Theory (Pustejovsky 1995, 1998).

The linguistic model is implemented in a generic representational framework, the GENELEX-PAROLE Entity/Relationship model, which provides a flexible lexicon architecture, an explicit descriptive language and a modular and non-redundant representation of information and a variable level of descriptive granularity. The descriptive objects relevant to each representational level, their structure, legal features and co-occurrence restrictions as well as their relationships are defined in an XML document type definition (DTD).

3 The Italian Lexicon

In the framework of the CLIPS project, the Italian PAROLE-SIMPLE lexicon was extended with the addition of a phonological level, the revision and refinement of existing entries and the extension of the lexical coverage. The PAROLE-SIMPLE-CLIPS lexical resource (henceforth PSC)[3] consists now of 387,250 phonological entries, 53,000 morphological entries, 64,500 syntactic entries and 57,000 semantic entries[4] of one-word verbs, nouns, adjectives, adverbs and grammatical words.

The following sections are devoted to outlining the nature and richness of information provided in the lexicon, with a particular focus on the semantic representation.

3.1 The four-level architecture

The model architecture consists of four structured and interconnected modules that enable to formalize the behaviour of Italian lexical units along the various levels of linguistic representation. The descriptive layers are mutually independent although their entries are interrelated by means of either one-to-one, one-to-many or many-to-one links.

---

[2] European Advisory Group for Language Engineering Standards.

[3] 'PSC' is not the acronym of the lexicon and is only used here for the sake of brevity.

[4] ILC is responsible for the linguistic model and guidelines as well as for the full encoding of the whole set of phonological and morphological units, for 37,500 syntactic units and 28,500 semantic units while Thamus, Italian Consortium for Multilingual Documentary Engineering, encoded 27,000 syntactic units and 28,500 semantic units.

At the phonological level, information is provided regarding stress position, vowel openness and consonant pronunciation. The morphological level accounts for the morphosyntactic category and subcategory of lexical units, and for their inflectional paradigm. At the syntactic level, the functional behaviour of lexical units with respect to the elements they subcategorized for is described. Finally, the semantic level provides a classification of word senses and encodes the essence of words' meaning, their interrelations and the semantic constraints they impose on their context.

3.2 Syntactic representation

A *syntactic unit* generally encodes a single syntactic behaviour of a morphological unit by describing both the inherent syntactic properties and restrictions of the headword and its contextual properties, if any. The syntactic context *lexically-governed* by the head is described in terms of optionality, syntactic function, syntagmatic realization, and any kind of constraints on each subcategorized element. Syntactic structures entering in systematic frame alternations are clustered in a *complex* syntactic entry and linked by means of the *frameset* mechanism, a representational device that enables to establish the correspondence between their respective frame positions.

3.3 Semantic representation

In PSC, the semantic description is grounded on a very rich and comprehensive lexical model which tackles mishandled issues in lexical semantics and addresses innovative aspects of semantic representation. Word senses, encoded as *semantic units*, are richly endowed with a wide range of fine-grained and structured information, expressed by semantic relations or features.

3.3.1 The SIMPLE ontology

The SIMPLE Ontology consists of 157 semantic types designed for the multilingual lexical encoding of concrete and abstract entities, properties and events. This multidimensional type system is based on hierarchical and non-hierarchical conceptual relations and accounts for the principle of orthogonal inheritance (Pustejovsky & Boguraev 1993). The ontology reflects the GL assumption that lexical items are multidimensional entities which present various degrees of internal complexity and thus call for a lexical semantic description able to represent different ranges of meaning components. The GL framework offers the formal means to uniformly represent this heterogeneous conceptual complexity of lexical meanings. Pustejovsky (1995: 61) defines in fact the semantics of a lexical item as a structure involving four different components. One of these, the *Qualia Structure,* which consists of four qualia roles,[5] enables to capture key orthogonal aspects of word senses. Accordingly, the ontology

[5] The *formal* role identifies an entity among others; the *constitutive* role expresses the entity's composition, its partonomic properties; the *agentive* role provides information about its coming about; the *telic* role specifies its function.

consists of *simple types* – which can be fully characterized in terms of one-dimensional taxonomic relations – and *unified types* – which also incorporate orthogonal meaning dimensions and thus require a multidimensional organization. Assigning a semantic type to a lexical unit does not simply mean ascribing it a mere semantic label but rather endowing it with a structured set of semantic information that is crucial to the type's definition. The relevance of each information piece entering in the definition of a semantic type is marked by a weight feature indicating whether it is 'type-defining' or 'optional'.

Besides the ontological typing, the information types encoded in a semantic entry range over domain of use, Aktionsart, derivational links, synonymic relations, membership in a class of regular polysemy and any other distinctive semantic features. Of particular interest, then, is the information encoded in the *Extended Qualia Structure* and the Predicative Representation.

3.3.2 The Extended Qualia Structure

In the framework of the SIMPLE model, the Qualia representational tool was extended by defining, for each of its four roles, a sub-hierarchy of values expressed in terms of semantic relations linking either intracategorial or cross-categorial semantic units. The design of the 60 Extended Qualia relations[6] allowed to express finer-grained distinctions both for structuring the componential aspect of word meanings and for capturing the nature of their relationships. They enabled to indicate not only the fact that an entity *has* a function, an origin and a composition but also the *type* of functionality ('used_for', 'used_against', 'used_as'); origin ('caused_by', 'created_by', 'derived_from') and internal constitution ('made_of', 'has_as_part', 'has_as_member') that characterizes it.

Besides the traditional paradigmatic lexical relationships of hyperonymy,[7] partonomy and synonymy, a qualia based semantic representation allows therefore the expression of new semantic links on the syntagmatic axis that inform on contextual relations and can therefore be considered as collocation relations, e.g.: *lettera, scrivere* (letter, write); *medico, curare* (doctor, cure).

Qualia structure shows less appropriate for formalizing meaning dimensions of abstract entities and events than for concrete entities. The difficulty to model the semantics of such items seems however imputable to the intrinsic complexity of their lexical semantics rather than to the inadequacy of the Qualia theory. Qualia roles have indeed been most helpful in providing the means to encode entities to which no 'isa' relation could sensibly be associated. Word senses such as *modo* (way), *causa* (cause), *scopo* (aim), which only convey a bare constitutive, agentive or telic dimension were in fact defined according to the meaning dimension they instantiate.

---

[6] http://www.ilc.cnr.it/clips/extended_qualia_structure.pps

[7] It is worth noting that the relation structuring the adjective class is, by contrast, the antonymic relation, just as in WordNet. Most of the time, a synonymic relation is also provided that allows to better understand the context of use of the adjective, e.g.: *alto / acuto* vs. *alto / grande* (high /acute; high / big).

3.3.3 The Predicative Representation

One of the crucial and innovative aspects of the PSC lexicon is the encoding of semantic frames and the connection between syntactic and semantic information. Each predicative entry is connected to a single lexical predicate. A predicate, by contrast may be either linked to a single entry or shared by all members of a derivational paradigm and related to each of them through an appropriate type of link. The argument structure is described in terms of predicate's arity, status (true or shadow), semantic role and semantic constraints of each semantic argument. The link between the syntactic and semantic levels is realized through the projection of the predicate-argument structure onto the syntactic subcategorization frame and the link of semantic arguments to the corresponding place holders in the syntactic frame.

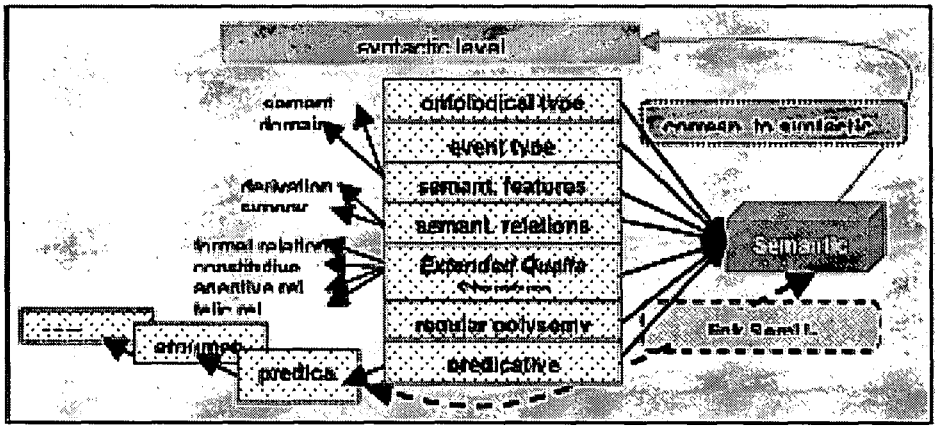

Figure 1. A semantic unit: information content

3.4 Template-driven encoding methodology

During the lexicon development, an innovative template-driven encoding methodology was adopted. Schematic structures containing clusters of structured, language-independent information corresponding to the semantic content of ontological types were proposed to the lexicographers in order to guide the encoding process. Such a methodology ensured intra- and inter-language encoding consistency, thus easing the reusability of data.

4 Final Remarks and Future work

In this paper we reported on a large scale, multi-layered and general-purpose lexical resource of Italian language based on a rich model that conjugates complexity and rigour with innovation in the representation of linguistic information, especially as far as lexical semantics is concerned.

Computational lexical resources and their data management tools provide the crucial infrastructure for HLT applications. Their exploitation is, however, often hampered by a too specialized development approach. To be portable, shareable and profitably reusable, resources must in fact meet the requirements of wide coverage, genericity, coherent structuring

of information, richness and explicitness of description and variability of descriptive granularity. PSC answers them all. Firstly, the annotation ranging over all levels of linguistic description provide NLP tools and applications with the whole range of information needed about the behaviour of lexical items. Secondly, the refined syntactic description, the explicit representation of the multidimensionality of word meaning and the account of relationships holding among lexical units, the encoding of semantic frames and the enforcement of semantic restrictions on the arguments, the link between syntax and semantics and the innovative methodology of template-driven encoding are all elements that highlight the resource's richness and high quality. Thirdly, the membership of PSC in a network of 12 European monolingual lexica sharing a conceptual and representational model that represents de facto standard for lexical representation confers to this resource an additional value. The inter-consistency of the 12 resources represent in fact an invaluable achievement in that it enables a flexible searching and extraction of data, their reusability in multilingual NLP/HLT applications and paves the way for a cross-language linking.

As far as semantics is concerned, the wealth of information encoded in the Italian lexicon, and in particular the expressive power of qualia-based representations, allows, among other, to build semantic networks, extract core sets of domain-specific information and acquire lexical collocations.

PSC lexical knowledge base addresses a wide range of potential users and applications such as Parsing, Natural Language Understanding, Information Extraction, Text Mining, Word Sense Disambiguation, and Question Answering. Moreover, the lexical resource lends itself to customization and extension. Partial knowledge, relevant to specific NLP application-dependent models can be derived from this repository of information, by mapping the application model from the generic one.

Up to now, the lexical resource has been used as a gold standard for the evaluation of Italian data in different EU projects, such as shallow parsing and knowledge extraction in the LE-SPARKLE project, reference syntactic lexicon for the IDEAL parser in the framework of the *Multilingual Summarization for the Internet* (MUSI) project as well as in various semantic tagging experiments. Moreover, the NSF-EU *International Standards for Language Engineering* (ISLE) project has used the SIMPLE semantic model as the basis for the creation of a general schema of multilingual lexical entry with a view to developing a standard representation framework for multilingual computational lexicons.

More recently, addressing the issue of knowledge transfer across languages and namely of developing new lexical resources from existing ones, a pilot study has been conducted that provided encouraging results about the feasibility of inducing a semantically annotated French lexicon from PSC. At present, an ILC ongoing project aims at the semi-automatic link of PSC and ItalWordNet with a view to merging the whole information into a common representation framework. In the near future, the database enrichment is foreseen with the dynamic integration of multiword expressions, collocations and named entities, through a process of automatic acquisition of linguistic knowledge from texts.

The lexicon, which has already undergone an internal content and consistency check, will now be validated by the European Language Resources Association (ELRA) and made publicly available.

References

Busa, F., Calzolari, N., Lenci, A. (2001), 'Generative Lexicon and the SIMPLE Model; Developing Semantic Resources for NLP', in Bouillon, P. and Busa, F. (eds.), *The Language of Word Meaning*, Cambridge University Press, pp. 333-349.

Calzolari, N., Lenci, A., Zampolli, A. SIMPLE 'Plurilingual Semantic Lexicons for Natural Language Processing', in Zampolli, A., Calzolai, N., Cignoni, L. (eds.), *Computational Linguistics in Pisa. Linguistica Computazionale*, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo I, pp. 323-352.

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

Lenci, A. et al. (2000), SIMPLE Linguistic Specifications, Deliverable D2. 1, ILC-CNR, Pisa.

Lenci, A., Bel, N., Busa, F., Calzolai, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000), SIMPLE: 'A General Framework for the Development of Multilingual Lexicons', in *International Journal of Lexicography*, Vol. 13, n° 4, Oxford University Press.

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

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

Pustejovsky, J. (1998), 'Specification of a Top Concept Lattice', Brandeis University, 1998.

Ruimy, N., Gola, E., Monachini, M. (2001), 'Lexicography Informs Lexical Semantics: the SIMPLE Experience', in Bouillon, P. and Busa, F. (eds.), *The Language of Word Meaning*, Cambridge University Press, pp. 350-362.

Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M., Calzolari, N., Zampolli, A. (2002), 'CLIPS, a Multi-level Italian Computational Lexicon', in *LREC 2002, Third International Conference on Language Resources and Evaluation Proceedings*, Vol. III, Las Palmas de Gran Canaria, pp. 792-799.

Ruimy, N., Monachini, M., Calzolari, N. (2003), 'Un lexique électronique multi-niveaux de l'italien', CIL XVII, *Proceedings of XVII International Congress of Linguists Prague*, Czech Republic, July 24-29. Prague, Matfyzpress.

Ruimy, N., Monachini, M., Gola, E., Calzolari, N., Del Fiorentino, M.C., Ulivieri, M., Rossi, S. (2003), 'A computational semantic lexicon of Italian: SIMPLE', in A. Zampolli, N. Calzolari, L. Cignoni, (eds.), *Computational Linguistics in Pisa. Linguistica Computazionale*, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo II, pp. 821-864.

Ruimy, N., Roventini, A. (2005), 'Towards the linking of two electronic lexical databases of Italian', in Zygmunt Veutulani (ed.), *L&T'05 – 2nd Language Technologies as a Challenge for Computer Science and Linguistics*, April 21-23, 2005, Poznan, Poland. Wydawnictwo Poznanskie Sp. z o.o., pp. 230-234.

Ruimy, N., Roventini, A. (2005), 'Towards the Linking of two Electronic Lexical Databases of Italian', in Vetulani, Z. (ed.), L&T'05, April 21-23, Poznan, Poland. Wydawnictwo Poznanskie Sp. z o.o. pp. 230-234.

Ruimy, N., Bouillon, P., Cartoni, B. (2005), 'Inferring a Semantically Annotated Generative French Lexicon from an Italian Lexical Resource', *GL'2005, Third International Workshop on Generative Approaches to the Lexicon*, Geneva.

Sanfilippo, A., Calzolari, N., Ananiadou, S., Gaizauskas, R., Saint-Dizier, P., Vossen, P. (eds.) (1999), 'Preliminary Recommendations on Lexical Semantic Encoding'. EAGLES LE3-4244 Final Report.