Clips, a Multi-level Italian Computational Lexicon: a Glimpse to Data

Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari, Antonio Zampolli

Istituto di Linguistica Computazionale, CNR, Pisa
Via Moruzzi, 1
56124 Pisa - ITALY
e-mail: nilda.ruimy@ilc.cnr.it

Abstract

CLIPS is a multi-layered Italian computational lexicon based on the PAROLE-SIMPLE model. In this paper we briefly recall the main characteristics of the model and devote our attention to issues emerging from the encoding of large quantities of data, especially in relation to those types of syntactic and semantic information specific to our lexicon and that reflect innovative features of the underlying model. At syntactic level, we show how alternating structures may be encoded in a linguistically more elegant way by using framesets. We illustrate the connection between syntactic and semantic information, and show how the SIMPLE Italian lexicon approach to predicate selection has been refined in CLIPS. At semantic level, we illustrate the richness of information types encoded in a word sense description and the way such a wealth of data can be exploited. We stress in particular the expressive power of the Extended Qualia Structure yet mentioning some of its problematic aspects. We show that queries on qualia relations allow to retrieve lexical collocates, to extract domain specific information, semantic networks, and help interpreting modifying PPs in complex nominals. Finally, we show that features, which cut across the type hierarchy, have a stronger expressive power with respect to semantic types in identifying selectional preferences.

1. Introduction

CLIPS[1] is a three-year Italian national project which started in 2000 and whose overall objective is to build core, generic, large scale and reusable textual and lexical resources. On the one hand, CLIPS aims at meeting a crucial requirement of speech understanding by creating a phonic archive based on a corpus of spoken Italian. On the other hand, a significant part of this project is devoted to developing, in a joint project between ILC and Thamus[2], a flexible knowledge base of lexical data annotated with information relevant to NLP applications and regarding the various levels of linguistic description. At the end of the project, the CLIPS lexical resource[3] will consist of 55,000 lemmas encoded at phonological, morphological and syntactic level and of 55,000 semantically encoded word senses. Out of this overall number, ILC is responsible for the treatment of 30,000 lemmas and 30,000 word senses. So far, 22,000 lemmas have been encoded at syntactic level and 21,000 word senses are assigned a semantic description.

The theoretical and representational model on which CLIPS is grounded has proved its validity in a three-phase EC program: the PAROLE and SIMPLE projects[4]. We will limit ourselves here to recall briefly the main characteristics of the PAROLE-SIMPLE model which has been presented at the previous LREC conferences[5].

The theoretical model which underlies the information encoded in the lexica is grounded on the EAGLES project recommendations[6], and on the extended GENELEX[7] model. The linguistic specifications get also inspiration from the results of EUROWORDNET[8], ACQUILEX and DELIS EC semantic projects. At semantic level, SIMPLE[9] implements and extends major aspects of Generative Lexicon (GL) theory[10]. PAROLE-SIMPLE linguistic guidelines are implemented in the GENELEX-PAROLE Entity/Relationship representational model which provides a flexible and modular lexicon architecture and an explicit descriptive language. The information encoded at the different descriptive levels is mutually independent, although the three layers are connected. Entries may be related by a one-to-one, one-to-many or many-to-one links. A morphological unit is linked to one or more syntactic units which share the same morphological properties. A syntactic unit (henceforth, SynU) , on the other hand, is associated with one or more semantic units (henceforth, SemU) depending on the number of meanings that a syntactic

[1] 'Corpora e Lessici di Italiano Parlato e Scritto'.
[2] Italian Consortium for Multilingual Documentary Engineering.
[3] The CLIPS homepage is under construction at http://www.ilc.cnr.it/
[4] The first step: the MLAP PP-PAROLE project dealt with the elaboration of the linguistic specifications. LE-PAROLE thus developed generic, multifunctional and re-usable harmonised written language resources for 12 European languages. The program third phase consisted in the LE-SIMPLE project which aimed at building wide-coverage and multipurpose computational semantic lexica linked to the morphological and syntactic ones elaborated during the previous phase.
[5] Ruimy *et al.* 98; Bel *et al.* 2000, Lenci *et al.*, 2000.
[6] Sanfilippo *et al.*, 1996 ; Sanfilippo *et al.*, 1999.
[7] GENEric LEXicon, EUREKA project.
[8] Ide N., Greenstein D., Vossen P. (eds.), 1998.
[9] SIMPLE Specification Group, 2000; Ruimy *et al.*, forthcoming.
[10] Pustejovsky, 1995, 1998.

entry conveys[11]. A SemU, in its turn, has access to the syntactic information of the entry(ies) it is linked with[12]. A complete entry is therefore a progression through the information layers. The level of descriptive granularity is variable: the model enables a very fine-grained description to be performed, but allows a more shallow one too, in so far as the information provided meets the model requirements. Lexical entries are represented in SGML in PAROLE and SIMPLE, while the CLIPS lexicon has adopted the XML format.

The twelve PAROLE and SIMPLE lexica share the approach to the conceptual and representational model, the core set of information encoded as well as the representation type. Such features enable their reusability for different application purposes and make them ready for multilingual linking. They are now being enlarged following the same principles at the national level, in different EC countries. In Italy, the extension of the PAROLE-SIMPLE lexicon is currently performed in the framework of the CLIPS project. The core body of data is being extended with a new set of lexical units selected from the PAROLE corpus according to frequency-based criteria and described, consistently with the existing data, at the various levels of linguistic description[13].

In this paper we devote our attention to issues emerging from the encoding of large quantities of data, especially in relation to those types of syntactic and semantic information specific to our lexicon and that reflect peculiar and innovative features of the underlying model. We point out the fact that some principles followed in PAROLE and SIMPLE have been progressively revised and tuned to new requirements imposed by the handling of more and more data. We also show how the richness of semantic information encoded in the lexical entries can be exploited.

2. Syntax

2.1. Syntax and Semantics correlation

The existence of a correlation between meaning and syntactic expression seems to be an uncontroversial fact. All along SIMPLE and CLIPS we got evidence that the relationship holding between a word's syntactic properties and lexical semantics operates in both directions.

On one hand, the encoding at syntactic level of diathesis alternations, for example, and the distinction into inchoative and causative readings of verbs revealed a common set of meaning components shared by members of each syntactic subclass, which led to the partition of these 'transition' predicates respectively into CHANGE_OF_STATE and CAUSE_CHANGE_OF_STATE semantic types, according to whether the causation was specified or not. This fact would suggest a relationship between the membership in a syntactic class and the sharing with all other class members of a certain lexical semantics representation. As a matter of fact, verbs apparently similar to alternating ones but which do not display the alternation, such as *tagliare* 'cut' reveal a different set of meaning components[14] and require a different lexical representation.

On the other hand, the addition of a semantic representation to words syntactically described evidenced a tendency of semantic type members to map onto syntactically coherent classes. Clustering lexical units into semantic classes on the basis of their meaning components has in fact revealed common syntactic properties of the class members which had eluded the syntactic encoding of isolated words, performed by different encoders. To give but an example, once lexical units such as *amore, commozione, ansia, inquietudine, sconforto* 'love, emotion, anxiety, dejection' were clustered under a unique ontological type, the fact that these class members could virtually share a common abstract semantic predicate, e.g. *PRED_FEELING* with two arguments filling the 'experiencer' and 'cause' semantic roles evidenced their sharing of a syntactic structure with both an *of*_ and a *for*_PP. It is clear, however, that a role does not always map coherently onto a syntactic expression all over a semantic class: in speech act verbs, for instance, the 'topic' role may be represented differently, both in terms of form and function, *discutere di / interrogare su / comunicare / un fatto*. Moreover, SemUs sharing a semantic class and an abstract predicate may not overtly instantiate all the roles of that predicate: let us think of transaction verbs, as *comprare* 'to buy' that may not express the notions of 'seller' and 'money'.

In spite of these last two points, we nevertheless advocated, in the CLIPS project, a semantic-driven approach to syntactic encoding. We are in fact convinced that, from a methodological viewpoint, an even coarse-grained and provisional semantic classification of lexical units may be of great help to encoders in performing a consistent description of their syntactic behaviour. Moreover, the semantic perspective helps relaxing the notion of argument structure to encapsulate the so-called 'adjuncts' which are crucial to the semantics of predicates.

2.2. Framesets

At syntactic level, every different syntactic behaviour of a morphological unit gives rise to a SynU. But while the distinction of each idiosyncratic syntactic structure of a lexical unit is a recommended practice, some types of regular information lend themselves to be formally

[11] This one-to-many link is instantiated in two different cases: (1) polysemy, e.g.: *cimice* has one syntactic entry which clearly gives rise to different SemUs ('bug' (insect); 'thumbtack'; 'bug' (electronics)); (2) the same entity is semantically described from different perspectives, e.g.: *libro* 'book' is encoded under two different semantic types: INFORMATION, to describe the book content and SEMIOTIC_ARTIFACT, for the physical object.

[12] The PAROLE-SIMPLE model also provides for multilingual links between SemUs. This aspect is currently being addressed in the EAGLES/ISLE project.

[13] The encoding process is performed using the CLIPS software tool for data management which allows importation, creation, browsing, editing and exportation of data as well. The handling of semantic information is based on the architecture of the SIMPLE tool.

[14] The meaning of *tagliare* involves an instrument and cannot be conceived without an agent. Besides, the verb's semantics implies a notion of contact and of motion.

represented in a linguistically more elegant and economically more convenient way. This is the case of regular and systematic alternations of syntactic structures shared by a consistent number of lexical units. The PAROLE model makes provision for relating information throughout the lexicon by means of the *frameset* descriptive device[15]. At representational level, the use of framesets, which allow to capture generalizations on deep syntactic relations shared by whole classes of lexical units, avoids a time-consuming and cumbersome enumeration of subcategorization frames. Framesets enable in fact, as shown below, to encode in a unique syntactic entry systematic frame alternations and to establish a relationship between the slot fillers of the alternating structures.

```
<FrameSet
id="FSERG2"
comment="link btw. transitive causative and pronominal inchoative"
example="ha rotto il vaso / il vaso si e' rotto"
descriptionl="t-xa ip-xepro">
 <Related>
  <RelElement1
    description="t-xa">
     <WayToPosition
       targetposition="1">
  <RelElement2
     description="ip-xepro">
      <WayToPosition
        targetposition="0">
</WayToPosition></RelElement2></Related></FrameSet>


<SynU
 id="SYNUdisperdereV"
naming="disperdere"
example="la polizia disperde i dimostranti;la folla si disperse"
description="t-xa"
descriptionl="ip-xepro"
framesetl="FSERG2"></SynU>
```

The phenomena treated in the CLIPS lexicon by means of the frameset device are those frame alternations wherein alternants are strongly linked to each other and do not imply a significant change of denotation with respect to each other, as e.g. decausativization, locative alternation, simple reciprocal alternation, symmetrical alternation and so on.

3. Linking Syntax and Semantics

The linkage of syntactic and semantic levels constitutes one of the most crucial aspects of the PAROLE/SIMPLE model. In CLIPS, the approach to predicate assignment adopted in the Italian SIMPLE lexicon has been revised and refined in two different ways (Ruimy *et al*. 2000, 2001a, 2001b).

The PAROLE/SIMPLE model foresees that, at semantic level, predicative entries are ascribed a predicative representation which consists in the assignment of a semantic predicate, the specification of the type of link the entry holds with it and the description of the arguments: 'arity', semantic role of each argument and selectional restrictions. For those entries, the relationship between syntactic and semantic information encompasses the connection of syntactic and semantic frames and the link of semantic arguments to syntactic positions[16]. In the Italian instantiation of SIMPLE lexicon, a predicative representation was assigned exclusively to (derived or simple) frame-bearing lexical units[17]. In CLIPS, we conferred more relevance to the semantic status of the predicate and to the relationship that connects it to its affiliates by linking all members of a derivational paradigm[18] (either frame-bearing or not) to a predicate provided selectional restrictions allow it through a set of appropriate links. By way of example, *viaggiatore* 'traveller' is linked to PRED_*viaggiare*, with i) the specification that the SemU absorbes the first argument of the predicate, and ii) an appropriate correspondence link capturing the fact that no argument of the predicate maps onto a syntactic frame position.

In CLIPS we also revised the mapping strategy by allowing a unique semantic predicate map onto different alternating syntactic structures of a lexical unit, through the design of a set of appropriate correspondence links. As shown in the previous point, causative-inchoative alternation is handled at syntactic level, in a unique complex syntactic entry, by a frameset which accounts for the two alternating structures. At semantic level, our type system allows to distinguish the two alternants with a different type assignment since specific types for 'change' and 'cause change' events exist. In SIMPLE, the two semantic entries pointed to different predicates: a one-place predicate for the inchoative reading and a two-place predicate for the causative reading. In the CLIPS lexicon, this approach has been revised in order to allow a unique semantic predicate account for the different surface realizations. Let us take the case of *migliorare* 'to improve'. On one hand, at syntactic level, the alternating structures have two possible realizations (transitive, for the causative reading; intransitive for the ergative reading) which are related by a frameset within a complex SynU. The frameset explicits also the link between the slots fillers of the different structures (P0 trans. = Ø intrans.; P1 trans. = P0 intrans.). On the other hand, at semantic level, two SemUs are created and assigned respectively the type CAUSE_CHANGE_OF_STATE and CHANGE_OF_STATE. Each of these two entries is linked to the two-place predicate PRED_*migliorare* (ARG0: agent, ARG1:patient) with a 'master' type of link indicating that the SemU is a privileged lexicalization of the predicate. In the link between syntactic and semantic levels, a

[15] The concept of frameset is part of the proposals made by EAGLES and introduced to the GENELEX model.

[16] Note that some arguments may not be linked to any syntactic position and some positions may have no corresponding arguments.

[17] Hence, a Ø-frame verbal nominalization such as *viaggiatore* 'traveller' was linked to *viaggiare* 'to travel' semantic entry through a derivational relation but not by means of the predicate PRED_*viaggiare*.

[18] Linked to the verbal predicate are now not only deverbal nouns but also nouns from which verbs derive, e.g.: *colpo* is related to PRED_*colpire* 'to hit'.

correspondence is established between the argument structure of the SemU and the appropriate structure of the SynU: hence the 'cause change'-typed SemU *migliorare* is related to the causative syntactic structure by means of a bivalent isomorphic relation holding between arguments and syntactic positions, while the 'change'-typed one is linked to the intransitive structure through a relation indicating that (ARG0:agent) does not map on any syntactic position while (ARG1:patient) maps on P0. The combination of the information provided by frameset, type of link and appropriate correspondence relations enables to explicit all kinds of connections (between predicate and SemU, between SemU and SynU, and between alternating structures of a SynU) and this allows to avoid the creation of two semantic predicates with the consequence of highlighting the strong relationship existing between the alternants and their similarity from a semantic point of view.

4. Semantics

Our objective here is to illustrate the richness of information types encoded in a word sense description and the way such a wealth of data can be exploited. In our lexicon, a semantic entry consists of a very rich bundle of information including type membership and its hierarchical position in the ontology, mapping to a different ontology[19], domain of use, gloss, type of event (for event-denoting entries), morphological derivation relation, logical polysemy class membership, synonymy (for adjectives), distinctive features and, as shown earlier, predicative representation and link to the corresponding syntactic entry. Besides, a substantial part of the information is encoded by means of the *Extended Qualia Structure*. We would like therefore to stress the advantages of qualia-based representations and the expressive power of this representation language, yet mentioning some of its problematic aspects. Beforehand, a very brief outline of the lexical semantics framework in which the encoding of data has been performed seems to be in order.

Following Generative Lexicon approach, the SIMPLE model[20] relies on the assumption that lexical units differ as to the degree of intricacy their semantics conveys. The GL theory enables to perform expressive and uniform lexical semantic representations of meanings of heterogeneous complexity. Pustejovsky defines in fact the semantics of a lexical item as a structure involving different components[21]. One of these, *Qualia structure* enables to express orthogonal aspects of word meaning whereas a unidimensional (even multiple) inheritance can only capture standard hyperonymic relations. As a matter of fact, a substantial amount of word senses denote a complex bundle of information and their meaning, which consists of orthogonal dimensions, cannot be exhaustively captured in terms of a mere subtype relation. An adequate description of their semantic content requires that all of the meaning dimensions be taken into account. Qualia structure allows to encode this multidimensionality by means of four Qualia roles which structure the information regarding essential aspects of a word's meaning. The *formal* role identifies an entity among others; the *constitutive* expresses the entity's composition; the *agentive* provides information about its coming about; the *telic* specifies its function.

4.1. The Extended Qualia Structure

In SIMPLE, Qualia structure has been modified to meet the requirements of the GENELEX model which imposed the implementation of Qualia roles in terms of relations between SemUs and of valued features. Qualia structure has been moreover made simultaneously richer and stricter. Richer, in that subtypes have been created by extending the set of possible values for each qualia role[22]. Stricter, in that this enlarged set of values allows to express finer-grained distinctions for describing adequately the relationships holding between so many senses. Qualia relations have also been marked as to their relevance in a type definition: either as 'type-defining', i.e. encoding information that intrinsically characterizes a semantic type, or 'optional', i.e. conveying non strictly essential — mainly world knowledge — information.

The whole set of 64 qualia relations devised has played a crucial role in defining the distinctive properties of SIMPLE ontology semantic types[23]. They have enabled to provide an exhaustive characterization of different levels of complexity of lexical meanings and to capture, besides the essence of a word denotation, additional meaning components that are important to a thorough lexical description. At encoding level, their adequacy for capturing key aspects of the lexical semantics of words, especially as far as nouns are concerned, results clearly from a parsing of traditional dictionary definitions: figures 1 and 2 below show that the meaning components which can be isolated in lexicographic definitions generally map quite easily on the dimensions expressed via qualia roles.

[19] LexiQuest semantic classes.

[20] See also Busa *et al.*, 2001; Calzolari *et al.*, forthcoming.

[21] Pustejovsky, 1995, 61.

[22] By way of example, the Telic role is expressible not only by a generic telic relation but also by more specific ones, such as 'object_of_the_activity', 'used_for', 'used_as', 'used_by', etc.

[23] The SIMPLE semantic type system, whose top types are mappable on the EuroWordNet ontology (Roventini *et al.*, this volume), has been slightly enlarged in CLIPS and consists now of a set of 157 language-independent semantic types, which are of two different kinds:

- simple (one-dimensional) types, fully characterizable in terms of a hyperonymic relation, e.g.: ROLE > HUMAN > LIVING_ENTITY;
- unified (multi-dimensional) types, only identifiable through the combination of a subtyping relation + the reference to orthogonal (telic or agentive) meaning dimensions, e.g.: CAUSE_ACT, unified type which inherits not only the properties of its supertype ACT but also an agentive meaning dimension.

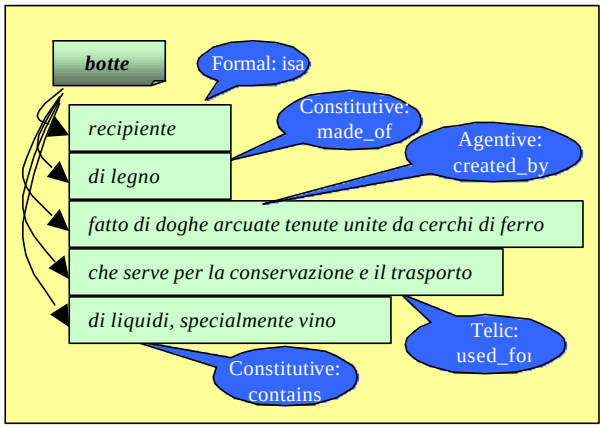

Figure 1: dictionary definition for *botte* [24]

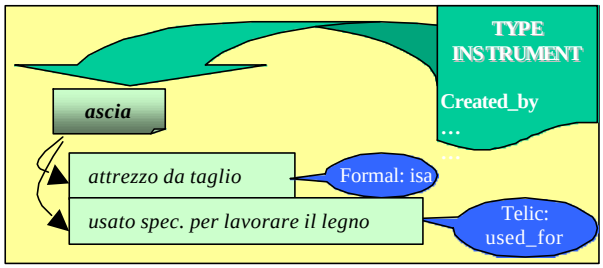

Figure 2: dictionary definition for *padella* [25]

Dimensions not explicitly expressed in the definition are inferrable by virtue of the word's membership to a semantic type. In figure 2, the missing agentive dimension is inherited by the SemU since the agentive relation 'created_by' is a type-defining information for INSTRUMENT type members.

Qualia are especially useful to get around the problem of assigning the genus term to some abstract entities not easily formalizable from a semantic point of view and which lexically instantiate only the Telic, Agentive or Constitutive quale. In other words, for those nouns for which a genus term can hardly be found e.g.: *scopo*, 'aim', qualia roles provide directly the interpretation, which is transparent in the structuring of the semantics.

Clearly, qualia relations present also problematic aspects. Their lexical instantiation, for instance, is not always straightforward and may pose challenges, especially as for Telic and Agentive roles. It may in fact be the case that a word meaning clearly convey a linguistically relevant dimension, and yet that its lexical expression be problematic. For the Telic role, this phenomenon may occur when:

- the denoted entity has an underspecified function, as happens most often for top elements of types, e.g.: *strumento*: *arnese, spec. tipico di un'arte o di un mestiere, che serve ad eseguire determinate operazioni*[26].

'Tool: device, esp. typical of an art or a profession, able to perform particular operations.'

- a unique word or MWU is inadequate to express the dimension, e.g.:

biblioteca: sala o edificio in cui sono raccolti e ordinati i libri destinati a consultazione, lettura, studio e sim.

'Library: room or building where are gathered and ordered books for consultation, reading, study and so on.'

- the functions are multiple, e.g.:

carta: materiale ottenuto da un impasto di sostanze fibrose, gener. cellulosa, che si presenta in fogli sottili ed è usato spec. per scrivere, imballare, ecc.

'Paper :Material obtained from cellulose fibers which is constituted by thin sheets, and is used especially to write, pack, etc.'

In some cases, a solution could be to allow a semantic class rather than a specific SemU as target of the relation; in most cases, however, this would not even be sufficient[27]. Similar problems emerge when dealing with the Agentive dimension felt in *dimenticare* 'to forget' or *morire* 'to die'. For all such cases, in order to preserve linguistically relevant information while avoiding underspecified and therefore non informative or odd relations, the problematic qualia relations are substituted for a lexically underspecified information expressed in terms of features.

All linguistic theories come up against difficulties: they are efficient in some areas, the ones of perceptible realities, natural species and manufacts but are less adequate to represent suitably the information in other domains. Qualia roles too seem to be less appropriate for capturing and formalizing meaning dimensions of abstract nominals and underspecified events than they are for concretes[28]. As a matter of fact, abstracts and events present a twofold aspect. On the one hand they are intrinsically complex: they are neither perceivable, nor measurable, and are not as easily understood as objects and concrete events; on the other hand the information about their semantics is rather subtle and with vague boundaries and the lexicographer's subjectivity, which plays a crucial role during a lexicon building process, affects even more the resulting analysis of those entities. The difficulty we faced in representing the meaning of such lexical units seems therefore to be imputable to the intrinsic complexity of their lexical semantics.

4.2. A glimpse to data

Looking at the lexical entries of our lexicon, some information could seem redundant at first glance. As a matter of fact, each one plays a different role, has a specific informative value and provides a different knowledge when combined with other information types. In the following points, we show that accessing qualia relations data via queries devised in the CLIPS tool, a wide range of information interesting for many application systems may be retrieved or inferred.

[24] 'Barrel: wooden container made of curved staves held together by metal strips used for keeping and transporting liquids, especially wine'.

[25] 'Hatchet: cutting tool used especially to work wood'.

[26] De Mauro, 2000.

[27] How can we express the 'instrument' relation for the SemU BODY_PART, INSTRUMENT, CONCRETE_ENTITY are far too wide since only a few members of these types may be used as instruments to hit.

[28] Ruimy *et al.*, 2001.

4.2.1. Lexical Collocates

Qualia relations enable to establish a connection between a word sense and a number of events or entities strictly related to its meaning and to define the role of those events/entities in the lexical semantics of the word itself. From such syntagmatically-related word pairs which express the lexical context of the entry, lexical collocates may be acquired. Let us take as example the semantic type CLOTHING whereby entries are subclassified according to different targets of the 'isa' relation, i.e.: (1) *indumento* 'cloth', (2) *calzatura* 'shoe', (3) *accessorio* 'accessory' , (4) *gioiello* 'jewel'. Each subclass member shares the same value for the telic relation 'object_of_the_activity', i.e. for subclass (1): *indossare;* (2): *calzare;* (3) and (4): *portare*[29]. Such word pairs allow to identify typical objects of the mentionned verbs, e.g.: *indossare un vestito, calzare gli stivali, portare una cravatta, una collana* and therefore to enforce adequate selectional restrictions on the predicates. In the SEMIOTIC_ARTIFACT type, by contrast, it is the agentive relation 'created_by' that allows identifying typical objects. In fact, *libro* 'book' and hyponyms are characterized by the targets *stampare* 'to print' and *rilegare* 'to bind' while SemUs as *manifesto* 'poster' have as unique target *stampare* and SemUs as *blocco* 'notebook' only *rilegare.* On the other hand, typical subjects may also be extracted, as for instance, for the verb *contenere* from the 'used_for' telic relation of members of CONTAINER and FURNITURE types, e.g.: *barile* 'barrel', *tanica*, 'tank', *vetrina* 'glass cupboard', *cassettone*, 'chest of drawers'. Typical subjects may also be retrieved through the constitutive relations i) 'constitutive_activity', which links animals to their typical activity, either typical movement or distinctive activity evidenced by corpus data, e.g.: SemU: *serpente* ® *strisciare* 'snake, to slither'*; pesce* ® *nuotare* 'fish, to swim', *zanzara* ® *pungere* 'mosquito, to sting', etc. or ii) 'typical_of' relating typical sounds to the animals producing them, e.g.: *gracidare* ® *rana* 'to croak, frog'*; frinire* ® *cicala* 'to creak, balm-cricket'.

4.2.2. Semantic Networks

The whole set of qualia relations in which a single keyword is involved throughout the lexicon enables to retrieve semantic networks[30]. For example, a query on the SemU *capra* 'goat' as target of all qualia relations it is used in enables to extract a set of 23 semantically-related words. The kind of relationship each retreived word holds with the keyword is explicitely provided by i) the qualia role and subtype of qualia relation it is used in and ii) the semantic type which indicates its location within the ontology. The closest words are obviously those sharing the keyword semantic type and whose relationship to it is further expressed by means of features indicating sex and age, e.g. *caprone* 'billy-goat' vs. *capretto* 'goatling'. The extracted word set consists of typical body parts, *vello* 'fleece'*;* typical location *ovile* 'fold'; typical activities the goat is agent or patient of, i.e. *belare* 'to bleat', *mungere* 'to milk' with respective nominalizations; 'products of': meat: *capra*, cheese: *caprino*, leather: *capra, marocchino*, wool: *angora, mohair, cashmere* as well as a metaphorical use for humans.

4.2.3. Domain Specific Information

Targets of qualia relations permit to capture orthogonal relationships existing between word senses all over the lexicon, regardless of their semantic type membership. Navigating through the database and searching alternatively by qualia relations and specific SemUs, information on particular domains can be extracted. Let us investigate, by way of illustration, the verb *mangiare* 'to eat' used in the Telic relations 'used_for' and _of_the_activity'. The 'used_for' relation allows to retrieve SemUs belonging to different areas of the ontology as *tavola* 'table' in the type FURNITURE; *posata* 'cutlery' in INSTRUMENT and, through it, *coltello, forchetta* 'knife, fork' etc.; in BUILDING the places where meals are served: *ristorante, trattoria,* etc. The 'object _of_the_activity' relation, on the other hand, enables to capture a large set of entries distributed over 5 different types, as for example: *arrosto* 'joint' in ARTIFACT_FOOD; *carne* 'meat': FOOD; *coniglio* 'rabbit': SUBSTANCE_FOOD; *mela* 'apple': FRUIT*; cavolo* 'cauliflower': VEGETABLE. This set is further enrichable with SemUs to which no such telic relation was assigned since they do not denote food that is properly eatable but rather that is used as ingredient. Such entries are retrievable via the feature PlusEdible, e.g. : *lievito* 'yeast': NATURAL_ SUBSTANCE*; alloro* 'laurel': FLAVOURING, etc. Moreover, the SemUs belonging to the type ARTIFACT_ FOOD allow to access, by means of the Agentive relation 'created_by', the different verbs denoting general or specific cooking processes: *cucinare, cuocere, arrostire, bollire, friggere,* 'cook, roast, boil, fry', etc. Investigating these verbs as targets of the relation 'used_for', entries belonging to different semantic types are captured. On the one hand, through the target *cuocere*, generic containers such as *padella* 'frying pan' and *pentola* 'pan' are retrievable, whereas through specific cooking verbs more sophisticated results can be obtained, e.g. 'used_for *friggere*' ® *friggitrice* 'fryer'*;* 'used_for *bollire*' ® *bollitore* 'kettle'*;* and even, combining telic and constitutive relation: ['used_for:*bollire*' + 'concerns: *pesce*' 'fish']® *pesciera* 'fish kettle'. On the other hand, through the target *cucinare,* cooking utensils such as *mestolo* 'ladle' etc. are extractable from the type INSTRUMENT[31]. The SemU *cucinare,* as target, this time, of the telic relation 'is_the_activity_of', enables to access the area of professions from which entries such as *chef, cuoco, cuciniere* 'chef, cook' can be extracted[32].

[29] Which all translate into English as 'to wear'.

[30] Similar experiments were performed in Acquilex (Calzolari, 1988) and by Fontenelle (2000) who applied an extension of the Mel'èuk lexical functions paradigm to a bilingual lexical-semantic database.

[31] Those with a more specific use, e.g. *grattugia* 'grater' are anyway retrievable through the domain 'Cuisine'.

[32] Again, other less specific professions from the same area, such as *sguattero, lavapiatti* 'scullery-boy' 'dishwasher' may be retrieved through the domain information.

4.2.4. Complex Nominals Disambiguation

N Prep N complex nominals are extensively used in Italian; the interpretation of their semantic structure is therefore crucial, especially for MT as well as IR and IE applications. It is a well-known fact, however, that disambiguating PPs of complex nominals is not an easy task[33] and that no generalization can even be made about the semantics of the preposition which may convey different meanings, as in: *duna di sabbia* 'sand dune', *bombola di gas* 'gas cylinder', *fetta di pane* 'slice of bread'. Accessing the lexical information of head nouns involved in some types of complex nominals and analyzing in particular their extended qualia structure, some clues to the interpretation of modifying PPs may be acquired. As a matter of fact, in the lexical representation of the head noun, qualia relations provide an interpretation of the modifier's semantic contribution not only in terms of kind of modification relation but also in terms of lexical specification. This is particularly the case for the telic role: e.g., in *canna da pesca* 'fishing rod', the PP disambiguation is made possible by querying the lexical representation of the head noun *canna* whereby the telic relation 'used_for: *pescare*' allows interpreting *da* as a telic marker. Similarly, complex nominals such as *macedonia di frutta*, 'fruit salad', and *succo di frutta* 'fruit juice' can be correctly interpreted and differentiated from each other by resorting, respectively, to the constitutive relation 'made_ of' and the agentive 'derived_ from' that are part of their head noun semantic representation[34].

Disambiguating on the basis of qualia relations offers the advantage of allowing some generalization over the PP interpretation according to the headword type information. Lexical units sharing a semantic type also share in fact qualia relations. Hence, headwords of complex nominals such as *canna da pesca*, *spazzolino da denti* 'tooth brush', *ferro da stiro* 'iron', *ago da cucito* 'needle', *mazza da golf* 'golf club', etc. which are all typed as instruments share a telic dimension expressed by the 'used_for' relation; their modifying PPs can therefore be univocally interpreted as denoting the headword's purpose.

A slightly different situation occurs with container-typed headwords such as *bicchiere, canestro, bombola* 'glass, basket, cylinder', etc. that enter in different complex nominals, i.e. *bicchiere di vetro, bicchiere di birra, canestro di vimini, canestro di frutta.* Here again, the PP semantics can be disambiguated by resorting to the head noun qualia structure, and, in this case, to the constitutive quale which informs, by means of appropriate relations, about both the material the entities are made of and their content. However, this time headword descriptions do not provide an explicit lexical specification: in the entry of '*bottiglia'*, for example, the 'contains' relation is filled by the generic SemU *'liquido'*, a SUBSTANCE-typed entry. It is therefore not directly but on the basis of type constraints that lexical units entering in complex nominals headed by '*bottiglia*'-like entries are interpreted as potential fillers of the 'contains' relation, thereby suggesting a 'content' rather than a 'material' interpretation of the modifying PP.

4.2.5. Selectional Restrictions

Another peculiar feature of our lexicon model is the enforcement of selectional restrictions on the arguments of semantic predicates. This is not a straightforward issue, however, since i) restrictions are to be taken as selectional *preferences* in prototypical, non metaphorical contexts, rather than as absolute constraints; ii) a balance must be struck between too fine-grained restrictions that may rule out possible combinations of lexical items and too loose ones that would turn out to be totally uninformative. In our model, restrictions are expressible in terms of semantic type, semantic class, SemU, features, or 'notions' combining these different expressive means. Although semantic types have indeed been used, and often successfully, to indicate preferences, e.g.: *indossare* 'to wear: [ARG0: HUMAN; ARG1: CLOTHING]; *dichiarare* 'to declare': [ARG0: HUMAN, HUMAN_GROUP; ARG1: EVENT]; they have in many cases proved to fail in capturing the full range of semantic arguments. For instance, while selecting the type HUMAN for the agent of 'eat' rules out animals; its supertype LIVING_ENTITY encompasses undesirable vegetals[35]. On the other hand, restricting this predicate's patient to the type FOOD filters out relevant candidate SemUs not encoded under the FOOD type hierarchy, such as vegetables and fruits.

In the CLIPS lexicon, the use of semantic features for marking predicate's arguments has therefore been preferred and, in this view, the feature assignment to SemUs has been extended with respect to SIMPLE. Features, which cut across the type hierarchy, allow in fact to capture larger sets of lexical units and are therefore deemed more suited to identify preferences. For the patient role of the predicate 'eat' [ARG1: +edible], the feature +edible enables, for example, to capture lexical units distributed over eleven different semantic types and to link the role lexical realization to an exhaustive list of relevant SemUs encoded in the lexicon.

5. Concluding Remarks

In this paper, we have presented some aspects of a multi-layered lexicon based on a model which has proved its validity in two important EC projects in the framework of which generic and large lexicons for all European languages were built. We have pointed out some aspects of the encoding strategy which have been revised in order to achieve a more elegant and coherent treatment of lexical units, in particular as far as the handling of predicative representation is concerned. We have shown how the multi-level nature of the lexicon enables to gain deeper insight into the type of context a lexical unit is inserted, especially as far as both syntactic and semantic restrictions are concerned. We have illustrated how the wealth of semantic information encoded in the lexical entries and particularly in the qualia structure can be

[33] Johnston and Busa, 1999.

[34] Note that, for the latter case, the agentive quale also accounts for the action the modifier undergoes, i.e..: *succo*: 'created_by: *spremere'* 'squeeze'.

[35] Such cases are solved by using 'notions' which combine information, e.g. for the agent of 'eat', the notion 'animate' which includes both HUMAN and ANIMAL type hierarchies.

exploited in many different ways.

Although largely used, the SIMPLE-CLIPS model — which now imposes itself as a *de facto* standard — is not crystallized but rather in continuous evolution and refinement. This aspect confers to CLIPS a status fairly different from the one of a mere implementation project: CLIPS does present innovative and challenging research aspects. For its richness and its multifunctional nature, the model lends itself to be enriched with the integration of further data such as multi-words units, collocates, information acquired from corpora, as well as a multilingual layer. It is currently being used in the framework of the NSF-EU ISLE project as the basis for the creation of multilingual lexical entries and this confers even more value to the integrated lexical suite it represents .

References

Bel, N., Busa F., Calzolari N., Gola E., Lenci A., 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 Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000), M. Gavrilidou et al. (eds.), Athens.

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. (1988). The Dictionary and the Thesaurus can be Combined, in M. Evens (ed.), Relational Models of the Lexicon, Cambridge, Cambridge University Press, pp. 75-86.

Calzolari, N., Lenci A., Zampolli A. (forthcoming). SIMPLE: Plurilingual Semantic Lexicons for Natural Language Processing, in Linguistica Computazionale, Giardini Editori, Pisa.

De Mauro, T. (2000) Il Dizionario della lingua Italiana, Paravia

Fontenelle, T. (2000). "Bilingual lexical database for Frame Semantics", in International Journal of Lexicography, Vol. 13, n°4, Oxford University Press.

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

Ide, N., Greenstein D., Vossen P. eds. (1998): Special Issue on EuroWordNet, Computers and the Humanities, Vol. 32, 2-3, 117-152.

Johnston, M., Busa F. (1999). Qualia Structure and the Compositional Interpretation of Compounds, In *Breadth and Depth of Semantic Lexicons*, E. Viegas (ed), Kluwer.

Lenci, A., Bel N., Busa F., Calzolari 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.

Levin, B. (1993). English Verb classes and Alternations, A Preliminary Investigation, The University of Chicago Press, Chicago.

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

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

Roventini, A., Ulivieri M., Calzolari N. (2002). Integrating Two Semantic Lexicons, SIMPLE and IWN. What Can We Gain?, this volume.

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 1st International Conference on Language Resources and Evaluation (LREC-1998), A. Rubio et al. (eds.), Granada.

Ruimy, N., Del Fiorentino M.C., Monachini M., Ulivieri M. (2000). Simple Lexicon Documentation for Italian, SIMPLE Project LE4-8346, WP03.9, final report.

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, 350-362.

Ruimy, N., Monachini M., Calzolari N. (2001a). Specifiche Linguistiche e Manuale di Codifica - Livello Sintattico, CLIPS-WP5, Pisa.

Ruimy, N., Monachini M., Calzolari N. (2001b). Specifiche Linguistiche e Manuale di Codifica - Livello Semantico, CLIPS-WP5, Pisa.

Ruimy, N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri M. (forthcoming): A Computational Semantic Lexicon of Italian: SIMPLE, in Linguistica Computazionale, Giardini Editori, Pisa.

Sanfilippo, A. et al. (1996). Preliminary Recommendations on Subcategorization, Pisa

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.

Simple Specification Group: Lenci, A., Busa F., Ruimy N., Gola E., Monachini M., Calzolari N., Zampolli A. et al. (2000), Linguistic Specifications, Simple WorkPackage 2, Deliverable D2.1.