FEDERICA BUSA, NICOLETTA CALZOLARI, ALESSANDRO LENCI AND JAMES PUSTEJOVSKY

BUILDING A SEMANTIC LEXICON: STRUCTURING AND GENERATING CONCEPTS

1. INTRODUCTION

One of the main challenges for computational lexical semantics is to bridge the gap between on the one hand theoretical research on the organization of the lexicon and on the formal representation of word meaning, and on the other hand the increasing need by natural language processing (NLP) systems of accessing large repositories of lexical knowledge. The latter, in fact, represents one of the most critical bottle-necks in the development of robust and efficient systems endowed with linguistic intelligence, given the time-consuming and difficult effort required by the construction and maintenance of lexical knowledge bases. This problem has recently been tackled by designing and developing (both at the national and at the international level) general lexical resources based on a common model, with the aim to provide an explicit and (possibly) standard representation of the linguistic content of lexical items at various levels of representation. Actually, their design represents an important challenge for the computational linguistics community, both from the theoretical and the applicative point of view. The difficulty of this enterprise is given by the inherently multi-purpose and domain-independent vocation of such lexical knowledge bases, since they are intended not to be restricted to specific terminological domains or application types, in order to ensure the maximum degree of portability and reusability. In fact, while application needs and domain features set natural constraints on the format of the lexicons, as well as on the type of the information they must contain, similar constraints are not available for the design of a lexicon which aims to provide general linguistic knowledge. In such cases it is crucial to provide satisfactory answers to issues like: What information a computational lexicon must contain? How to represent this information? What is the format to give to a lexical entry?

In this chapter we illustrate the general model for semantic lexicons developed in the EU-sponsored SIMPLE project,[1] which involves the construction

[1] SIMPLE stands for *Semantic Information for Multipurpose Plurilingual Lex-*

H. Bunt et al. (eds.), Computing Meaning, Volume 2, 29–51.

of wide-coverage lexicons for twelve languages,[2] adding a semantic layer to the PAROLE syntactic lexicons. Even though SIMPLE is a lexicon building project, it has also addressed challenging research issues in the realm of lexical semantics grounded on, and connected to, a syntactic foundation. In particular, SIMPLE has tried to answer some of the issues above by largely grounding the architecture of the lexicon on the theoretical approach provided by the Generative Lexicon (henceforth GL), (Pustejovsky, 1995; Pustejovsky, 1998), where *qualia structure* represents the basic building blocks for structuring and generating the concepts expressed by word senses. This way, SIMPLE has benefited from the natural vocation of the GL to deal with the complex architecture of the lexicon, as revealed in both meaning variation and systematic polysemy, thus providing a consistent guidance for the construction, maintenance and customization of lexical entries.

The nature of the questions addressed in this chapter have been shaped by the requirements of SIMPLE: the size of the lexicon to be built, the multilingual framework, the difficulty of establishing a priori the granularity of the semantic representations for all lexical items, the necessity to provide the lexicographers with guidelines for developing consistent representations among all the languages and, finally, the requirement of flexibility and openness towards future extensions geared at particular applications.

The development of such a resource tackles questions that are at the core of lexical semantics research and SIMPLE represents an important opportunity for testing the viability of computational semantics models of lexical knowledge on a large scale and from a multilingual perspective. In particular, this latter aspect requires a careful identification of lexical conceptual structures to be shared by different languages.

More specifically, the following questions are at the core of the SIMPLE model:

A. what is the structure of the concepts expressed by lexical items?
B. how do these concepts differ in terms of their complexity?
C. how is this complexity characterized and by which formal means is represented?
D. do all major categories of the language involve the same conceptual representation?

In the remainder of this chapter, we first discuss some of the standard approaches to computational lexicon design and semantic classification. These

icons, sponsored by DG-XIII of EU, within the LE - Language Engineering Programme, and coordinated by Antonio Zampolli.

[2] Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish.

models are then briefly compared to the SIMPLE architecture for lexical entries, which is largely based on the notion of *extended qualia structure*, providing the basic syntax for recursively generating structured conceptual representations.

2. Semantic Types and Lexicon Architecture

Ontologies surely represent a key ingredient in lexical knowledge management, and actually this is the main reason for their renewed fortune in computational lexical semantics. Representing one of the meanings of a word minimally implies (i.) distinguishing it by other senses the same word might have, (ii.) capturing certain inferences which can be performed from it, and (iii.) representing its similarity relations with the meanings of other words. To this purpose, ontologies are powerful formal representational tools exactly because word meanings can actually be regarded as entities to be classified in terms of the ontology types. In this perspective, a given sense can be described by assigning it to a particular type, so that the ontology structure will account for entailments between senses, and similarities between word senses will correspond to the sharing of the same ontology type.

However, as models for the lexicon, ontology design must face an incredibly hard and challenging task, due to the difficulties and complexity of lexical knowledge. This is, in fact, inherently heterogeneous and implicitly structured. Moreover, polysemy is a widespread and pervasive feature affecting the organization of the lexicon. Finally, word senses are multidimensional entities which can hardly be analysed in terms of unique assignments to points in the ontology. 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 the lexicon and its multifaceted behaviour, which makes it closer to a kaleidoscope of senses, continually changing their relations and nature depending on the vantage point from which they are observed.

Research in cognitive psychology and lexical semantics has also shown that words crucially differ for the relative salience of different dimensions. For instance, while natural kind terms are mainly organized in terms of taxonomical hierarchies, a proper description of artifactual terms calls for the specification of their function (Keil, 1989). Similarly, different aspects of meaning are to be taken into consideration to provide a suitable representation of the content of abstract terms, verbs, adjectives, etc. Natural language complexity, thus, prevents the adoption of off-the-shelf type systems, and calls for

the design of architectures specifically tailored to capture the organization of the lexicon.

In particular, most of the current ontologies are affected by the so-called problem of *ISA–overloading* (Guarino, 1998). The prominent role assigned to the taxonomical ISA relation in the organization of the type system, in fact, lies at the base of important inefficiencies in the representation of word content in crucial areas of the lexicon. The current methodology for building ontologies is mostly centered around the question: *What is a certain entity?* This way, type systems fail to provide efficient representational tools for those word senses which cannot be satisfactorily classified in terms of this semantic dimension. Take for instance the case of words like *priority*, *materials*, *product*, *goal*, *target*, *link*, *mistake*, *dimension*, *member*, etc. The examples below show that the type of entities these nouns express is highly context-dependent, and in many cases very hard to identify too:

(1) a. Factories seemed to be China's highest priority. (*artifact*)

b. Getting food was his main priority. (*eventuality*)

c. The government has changed his priorities. (*abstract_entity*)

(2) a. IBM lauched a new product. (*software*)

b. The product of his best thinking is in the third chapter. (*abstract_entity*)

c. The bombing of Iraq was the product of their diplomatic efforts. (*event*)

(3) John made many mistakes

a. in his paper. (*typos*)

b. in his relationship with Mary. (*eventualities*)

(4) a. In the library I found many materials about the D-Day. (*semiotic_artifacts*)

b. The materials for the roof are in the garden. (*physical_objects*)

c. Please, send me all the *materials* you have about the new project. (*information*)

For instance, something is a *priority* if it is regarded to have precedence over other entities, independently of its specific nature. Similarly, the meaning conveyed by *materials* is exclusively functional, irrespectively of the specific 'substratum' (concrete or abstract) of the entity. Equally functional is the interpretation of *target*: in fact, an entity is a *target* if it fulfills a certain function in a given context. Similarly, anything can be a *link* as long as it connects two entities in a certain way, the specific way, however, being only determined by knowing what those entities are (cf. for instance the semantic difference between the noun phrases *the link between the webpages* and *the link between Rome and Milan*). The essential fact is that trying to characterize the examples in (1)-(4) by multiplying the senses of these nouns according to the different ISA dimensions is not a viable solution, apart from being totally unexplicative. In fact, such nouns can acquire new taxonomical interpretations depending on the linguistic contexts, and are core cases of the generative nature of the lexicon, thus providing important clues of the orthogonal dimensions that intervene in organizing and shaping the conceptual content of lexical items.

Representing such word senses in terms of type systems that rely too much or exclusively on the ISA dimension ends up in lexical characterizations which are often not fully adequate. One interesting example is provided by WordNet, where semantic description is provided by a 'verticalized' taxonomic hierarchy connecting a given synset to a top node. Thus, the backbone of the hierarchy (at least for nouns) is represented by the ISA relation. WordNet, notwithstanding its impressive capacity of structuring the lexicon, fails to offer satisfactory representations for nouns like the ones above, as the characterization of the senses of the noun *part* in Fig. 1 shows.

Notice that a twofold distinction is made: first of all, between *part* as a relation and *part* as an entity, and then between *part* as a concrete, physical object (e.g. a part of a car) and *part* as a psychological feature (e.g. a part of a theory). The problem is that neither of these distinctions is really justified, let alone it justifies the splitting of senses. In fact, a *part* has an inherently relational meaning, and being a part is not a matter of being concrete or abstract, but just of having a certain relation with something else. It is the nature of the entity to which something belongs as a part to determine whether it is abstract or concrete.

Verbs also raise problems for traditional monodimensional analyses. Consider for instance the case of (5). The verbs in (5) show an important analogy with the nouns in (1)-(4). The predicate is so underspecified that under an approach that aims at identifying each sense of a verb in a context with an a priori semantic class – as in (Levin, 1993) – the verb *keep* would not

http://www.springer.com/978-1-4020-0175-8

Computing Meaning
Volume 2
Bunt, H.; Muskens, R.; Thijsse, E. (Eds.)
2001, VI, 306 p., Hardcover
ISBN: 978-1-4020-0175-8