Part I

Fundamental aspects

1 Ontology and the lexicon: a multidisciplinary perspective

Laurent Prévot, Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, and Alessandro Oltramari

1.1 Situating ontologies and lexical resources

The topics covered by this volume have been approached from several angles and used in various applicative frameworks. It is therefore not surprising that terminological issues arise when the various contributions to the domain are brought together. This volume aims to create synergy among the different approaches and applicative frameworks presented.

Ontologies[1] are commonly defined as specifications of shared conceptualizations (adapted from Gruber, 1995 and Guarino, 1998b). Intuitively, the *conceptualization* is the relevant informal knowledge one can extract and generalize from experience, observation, or introspection. The specification is the encoding of this knowledge in a representation language (See Figure 1.1, adapted from Guarino, 1998b).

At a coarse-grained level, this definition holds for both traditional ontologies and lexicons if one is willing to accept that a lexicon is something like the linguistic knowledge one can extract from linguistic experience. However, a crucial characteristic of a lexicon is that it is linguistically encoded into words. In order to understand more subtle differences one has to look closer at the central elements of ontology creation: *conceptualization* and *specification*. What distinguishes lexicons and ontologies lies in a sharper interpretation of these notions.

Ontologies and semantic lexical resources are apparently similar enough to be used sometimes interchangeably or combined into merged resources. However, lexicons are not really ontologies (Hirst, 2004 and Chapters 12, 13). For example, synonymy and near-synonymy are very important relations for semantic lexicons, while there is no room for them in formal ontologies where concepts should be unambiguous and where synonymic terms are

[1] We follow here the accepted differentiation between Ontology (the philosophical field) and ontologies (the knowledge representation artefacts).

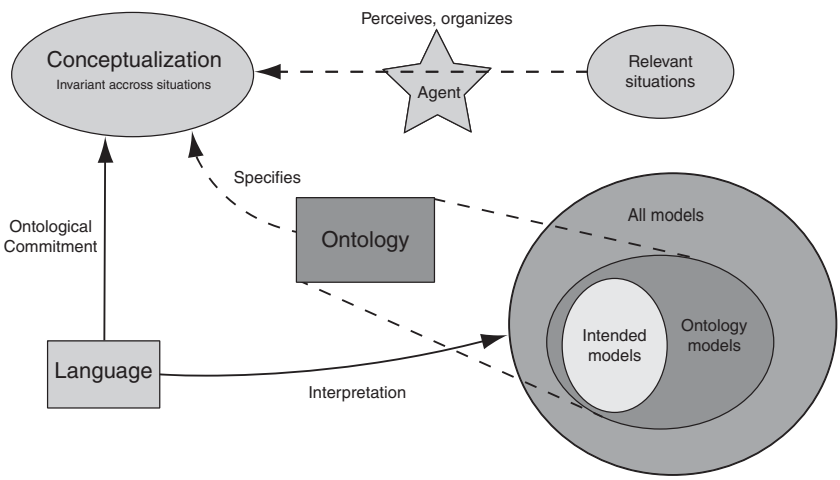

Figure 1.1 Conceptualization, specification and ontology

grouped under the same concept. From the ontological viewpoint the issue of synonymy is external and transparent to the ontological representation. Ontological discussions take place once synonymy issues have been resolved. Another example is the information about word usage (e.g., register) offered by lexicons but not relevant for traditional ontologies. Overall, linguistic resources, such as lexicons, are made of the linguistic expressions and not of their underlying concepts, while linguistic ontologies contain such underlying concepts.

The knowledge these resources attempt to capture has a very different nature, and in order to improve the management of the so-called *ontolex interface* it is useful to consider in some detail their differences, as we will see in the following subsections.

More practically, the important distinction we make in this volume is the supposed difference between formal and linguistic ontologies. According to the traditional view, formal ontologies are logically captured and formally well-formed conceptual structures, while linguistic ontologies are grounded on human language and are therefore 'linguistically conventionalized', hence often not formally precise, conceptual structures. The formal/linguistic opposition hides a much richer and layered classification that can be unveiled by sharpening the analysis of the resources in terms of conceptualization and specification.

At a terminological level, computational lexicons, lexical resources or relational lexicons differ from each other in a non-trivial way. However,

since this book deals specially with natural language processing (NLP) and Semantic Web issues, the lexical resources we consider are machine-readable and are therefore synonymous with *computational lexicons*. Finally, since relations are essential components of computational lexicons, we also take *relational lexicon* as a synonym in the context of this book.

The interface between ontology and lexicon (the ontolex interface hereafter) is born out of their distinct yet related characteristics. A lexicon is about words, an ontology about concepts, yet they both represent shared conceptualization, from the perspective of conventionalization. For applications in human-language technology, a lexicon establishes the interface between human agents and knowledge. For applications in the Semantic Web (Berners-Lee *et al.*, 2001), an ontology enables the machine to process knowledge directly. It is in this context that the ontolex interface becomes a crucial research topic connecting human knowledge to web knowledge.

1.1.1 Conceptualization

The nature of a conceptualization greatly depends on how it emerged or how it was created. Conceptualization is the process that leads to the extraction and generalization of relevant information from one's experience. A conceptualization is the relevant information itself. A conceptualization is independent from specific situations or representation languages, since it is not about representation yet. In the context of this book, we consider that conceptualization is accessible after a specification step; more cognitive-oriented studies, however, attempt at characterizing directly the conceptualizations (Schalley and Zaefferer, 2006). Every conceptualization is bound to a single agent, namely it is a mental product which stands for the view of the world adopted by that agent; it is by means of ontologies, which are language-specifications[2] of those mental products, that heterogeneous agents (humans, artificial or hybrid) can assess whether a given conceptualization is shared or not and choose whether it is worthwhile to negotiate meaning or not. The exclusive entryway to concepts is by language; if the layperson normally uses natural language, societies of hybrid agents composed by computers, robots, and humans, need a formal machine-understandable language.

To be useful, a conceptualization has to be shared among agents, such as humans, even if their agreement is only implicit. In other words, the conceptualization that natural language represents is a collective process, not

[2] Language here is no more than a representational formalism and vocabulary, and therefore is not necessary a natural language, but could be, for example, a predicate logic and a set of predicates and relations constituting the vocabulary of the theory.

an individual one. The information content is defined by the collectivity of speakers.

Philosophers of language consider primarily linguistic data and introspection for drawing generalizations to be used as conceptualizations for building natural language ontology. Traditional lexical semanticists will use mainly lexical resources as a ground for the conceptualization. Cognitive scientists might broaden the range of information sources, possibly including other perceptual modes such as visual or tactile information (see Section 1.3.1).

In our understanding, this is how a *linguistic ontology* is distinguished from a *conceptual ontology* that does not restrict its information sources to language resources. These kinds of ontology that acknowledge the importance of the agent conceptualization are called *descriptive* ontologies and they are opposed to *revisionary* ones (Strawson, 1959). A *descriptive ontology* recognizes natural language and common sense as important sources for ontological knowledge and analysis, while *revisionary ontology* refutes this position and is committed to capture the intrinsic nature of a domain, independently from the conceptualizing agents (see Masolo *et al.*, 2003; Guarino, 1998b, and Section 1.3.2).

In *lexical ontologies*, conceptualization is based on linguistic criteria, more precisely information found in lexical resources such as dictionaries or thesauruses. In many cases they are slightly hybrid since they feature mainly linguistic knowledge but include in many places world knowledge (also called encyclopedic or common-sense knowledge). Lexical ontologies are interesting because of the special status of the lexicon in human cognition. Indeed there are two notions of lexicon. A lexicon can be defined as a collection of linguistically conventionalized concepts, but in a more cognitive framework it is a store of personal knowledge which can be easily retrieved with lexical cues. In the context of this volume, we focus on the former definition of *lexicon*.

Engineering and application ontologies that have conceptualization grounded in shared experiences among experts are also relevant in the NLP context. How such ontologies can be integrated with more generic ontologies is of great interest in this volume (see Chapters 13 and 17, which explicitly deal with this issue).

Finally, a further refinement is introduced between linguistic conceptualizations derived from one unique language (monolingual linguistic ontology) or from several languages (multilingual linguistic ontology). Although language-based, the further generalization obtained through crosslinguistic consideration renders the conceptualization less dependent on surface idiosyncrasies. The issue is then to determine whether the conceptualizations based on different languages are compatible and, if not, how to handle them. Multilingual issues are extremely important for obvious applicative purposes, but their development might also help to investigate the complex relationship between language,

culture, and thought. A recurrent question for both cognitive science/NLP is the existence/need of a distinction between the so-called conceptual level (supposedly language independent) and the semantic level that would be deeply influenced by the language. These issues will be developed further both in the sections devoted to cognitive approaches (Section 1.3.1) and to NLP applications (Section 1.4.3).

The conceptualization process is a crucial preliminary for ontology construction. However, it is not the focus of this book and we encourage the reader to consult the more cognitive oriented contribution made in Schalley and Zaefferer, 2006.

1.1.2 Specification

The second operation is specification, as an ontology specifies conceptualization in a representation language. Apart from the level of complexity and explicitness, what is crucial is that ontologies, as language-dependent[3] specifications of conceptualizations, are the basis of communication, the bridge across which common understanding is established.

The nature of this language leads to the second main source of differentiation for ontologies. *Formal ontologies* are expressed in a formal language, 'informal ontologies' are, for example, expressed in natural language, and *semi-formal ontologies* combine both.[4] An important aspect of this distinction is the exclusion of ambiguity from formal ontologies while it is ubiquitous in semi-formal ones. However, this cannot be a blind generalization. Ontologies may be extremely rigorous and precise although formulated in natural language, and formality alone does not ensure rigour and precision.

Linguistic ontologies use the word senses defined in lexical resources (either informally or semi-formally as in WordNet) to create the concepts that will constitute the linguistic ontology. This move is a difficult one and if not performed carefully can lead to poor resources from an ontological viewpoint (see Chapter 3 for details on this problem). Still, in principle, nothing prevents a linguistic ontology from being formal.[5] It is the difficulty of such a project that makes linguistic ontologies only 'semi-formal'. More precisely

[3] Language-dependent does not mean here dependent to any given natural language but to the language used to formulate the ontology.

[4] Etymologically the 'formal' of 'formal ontology' also comes from the idea of not focusing on one area of knowledge but on principles equally applicable to all areas of knowledge. As such they operate at the level of the form rather than of the content. However, the more straightforward aspect of formality versus informality is emphasized here.

[5] Moreover, it is important to make the distinction between a linguistic ontology and an ontology of linguistics. The latter is an ontology concerning objects for linguistic description such as GOLD, Generic Ontology for Linguistic Description (Farrar and Langendoen, 2003). See GOLD web page (http://www.linguistics-ontology.org/gold.html) for more information.

axiomatizing the definitions (including the disambiguation of their terms) is still more of a research topic than a standard procedure for obtaining formal ontologies (see, however, Harabagiu *et al.*, 1999 and Chapter 3).

1.1.3 Scope

Three different levels of specificity for ontologies are recognized in ontological research and practice: upper-level, core (or reference), and domain ontologies. Foundational resources are sometimes confused with upper-level resources. They both concern the most general categories and relations which constitute the upper level of knowledge taxonomies. Foundational resources are further distinguished from upper-level by the additional requirement of providing a rich characterization, while upper-level resources include, for instance, simple taxonomies. They contrast with resources such as specialized lexicons or domain ontologies dealing with a specific domain of application that can be extremely restricted. The distance between upper and domain levels made it necessary to have an intermediate level: the *core* resources (see Figure 1.2). Core resources constitute the level at which is found intermediate concepts and links between foundational and domain resources. They can, however, vary greatly in content according to their main function: to provide a more specific but sound middle level or simply provide the mapping between the two levels. For example, MILO (MId-Level Ontology) is designed specifically to serve as the interface between upper and domain ontologies. Such mid-level ontologies can be considered as an extension of the upper ontology in the sense that they are supposed to be shared or linked to all domains. On the other hand, they also overlap greatly with a global resource since most of the terms at this level are

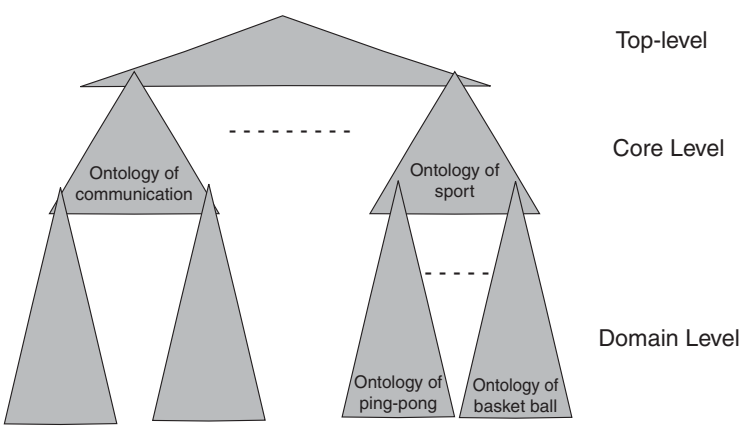

Figure 1.2 Scope of ontologies

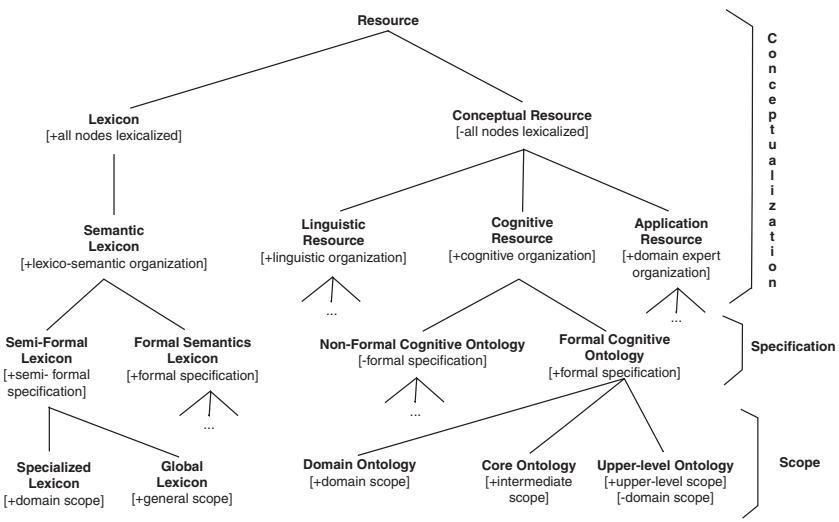

Figure 1.3 Ontolex resources taxonomy overview

linguistically realized, in comparison to many abstract and non-realized terms in upper ontology.[6]

More discussion on this issue is provided in the introductory Chapter 10 and in Chapter 13, where the notion of *global ontologies* is introduced for resources like WordNet, covering a broad scope while providing a good coverage by gathering all the entries a general purpose thesaurus could provide. Among traditional ontologies, CYC (Reed and Lenat, 2002) is also an example of global resource.

1.1.4 The ontolex interface

The previous sections allowed us to identify lexical resources and ontologies as objects of partially similar nature but differing with regard to their conceptualization, specification, and scope as illustrated in the taxonomy of Figure 1.3. These differences come from different research traditions. Ontologies and lexical resources, in their modern technical sense, historically belong to different applicative programs that have only recently been considered simultaneously.

[6] Note that the most recent version of IEEE upper ontology (www.ontologyportal.org) merged the original SUMO and MILO. Hence the distinction between upper and middle ontologies is blurred in this resource but the interface between upper ontology and lexicon is enhanced.

From an ontological viewpoint, the basic building blocks of ontologies are concepts and relations. Identifying these objects and deciding about their nature is a fundamental task of ontological analysis. A similar concern centred around terms and relations is found in lexical resources. These resources have sometimes been called *relational lexicons* (Evens, 1988) since the network of relations is supposed to contribute significantly to the meaning of the lexical entries. Concepts (or words) and relations are therefore the first two objects to consider while working with ontologies and lexical resources. This parallelism in their structure defines the ontolex interface.

Ontological analysis and construction handle concepts (for which words may or may not be available) that are grounded on knowledge representation arguments (homogeneity, clarity, compactness, etc.). On the other hand, lexical ontologies start from an existing and usually large vocabulary and come up with a sensible and useful organization for these terms. The work situated at the ontolex interface has therefore to find the best integration of both approaches. The exact combination of the conceptual information found in traditional ontologies and the lexical information is indeed the topic of most chapters of Parts III and IV of the present volume.

The ontolex interface also turns out to be extremely important in the design of multilingual resources. In the spirit of EuroWordNet (Vossen, 1998), these resources are typically constituted of several language-dependent monolingual resources mapped to an interlingua. Although this interlingua is generally unstructured (Vossen, 1998), giving it a structure is an important track of improvement followed in this domain (Hovy and Nirenburg, 1992) (see also Chapter 15). This structured interlingua might correspond to the conceptual level mentioned before. In addition to hold promise for language-engineering applications, this type of multilingual resource should facilitate the research on lexical universals and may also contribute to the recurrent universalists/relativists debate.

1.2 The content of ontologies

1.2.1 Concepts and terms

In an ontology, the nodes are of a conceptual nature and are called concepts, types, categories, or properties (see Guarino and Welty, 2000a). They are often characterized extensionally in terms of *classes* and correspond in this case to sets of instances or individuals. In ontologies directly derived from lexical resources, individuals (denoted by proper names and other named entities) are sometimes treated like other concepts. In some of these resources little attention has been given to the difference between classes and instances: they are both concept nodes of the resources and are represented in the same way.

Both classes and instances were entering in the same relation leading to the well-known **is-a** overload issue (see Chapter 3 for a detailed discussion of this issue). For example, until WordNet 2.0, each American president (e.g *Kennedy*) was given as a hyponym of *president*. Version 2.1 of WordNet added an **instance-of** relation for these cases. From a sound ontological perspective, a strong emphasis is put on the need for a clear distinction between these two components as made explicit by the distinction of an onomasticon, storing factual data, as a separate component of the Ontological Semantics (OntoSem) apparatus presented in Chapter 7 (see also Chapters 2 and 3).

The difference between a term-based lexicon and a concept-based lexicon is clear cut. However, the sense-based lexicon complicates the picture. In a sense-based lexicon like WordNet, the nodes of the resources are neither simple terms nor pure conceptual entities but word senses that correspond to a conventionalized use of a word, possibly coming from corpus-attested examples.[7] In WordNet, the nodes are *synsets*, i.e. sets of word senses that define sets of synonyms as made explicit in Chapter 2. Therefore, WordNet is primarily a lexicon since all its entries are linguistic expressions, but semantic structure defined by the synsets and their relations have frequently been used as a linguistic ontology (see Chapters 2 and 3 for issues with regard to this topic). The necessity of this intermediate semantic level is also discussed with more details in Chapters 14, 12 and 15.

1.2.1.1 The top-down approach to word senses In formal ontologies, ambiguity of terms has to be resolved as much as possible before entering the formal specification phase. The objective is to reach high precision for the intended meaning of each term in order to avoid misunderstandings. A central task of ontology building is to track down and get rid of ambiguities from the knowledge domain and to build more precise and reliable formal ontologies through analysis. An essential step of the ontological analysis process consists in determining a backbone taxonomy that provides the main categories and their taxonomic architecture organized along an **is-a-kind-of** relation. The top level of this backbone introduces, for example, the distinctions between objects, processes and qualities, between artefact and natural objects. Applying these structures to lexicons constitutes a 'top-down' approach to word senses since they will be strongly determined by the position of their attachment in the taxonomy. This approach is exemplified in Chapters 2 and 3.

1.2.1.2 The bottom-up approach to word senses In spite of its usefulness for knowledge representation, the top-down approach meets its limit when focus is put on natural language. Languages have productive mechanisms

[7] In Fellbaum, 1998: 24, synsets are described as lexicalized concepts.