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).

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

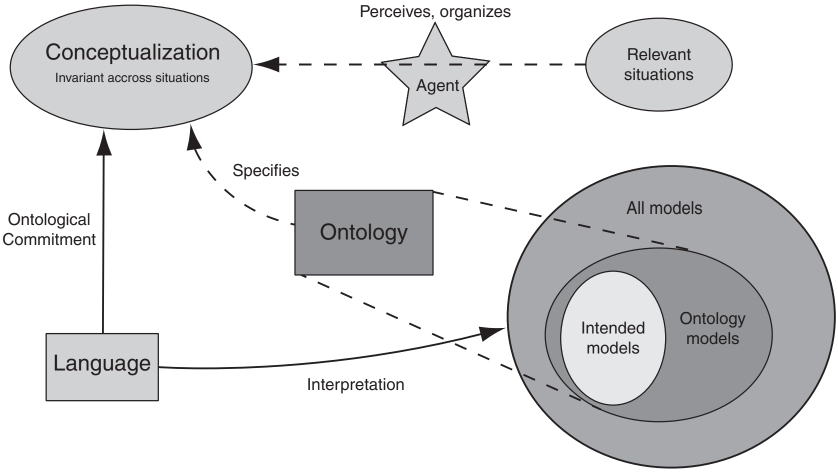

Figure 1.1 Conceptualization, specification and ontology

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,

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

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.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

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,

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

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.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

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

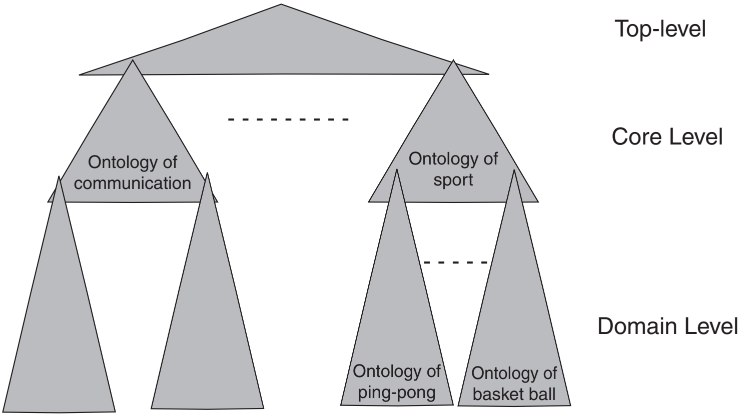

Figure 1.2 Scope of ontologies

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

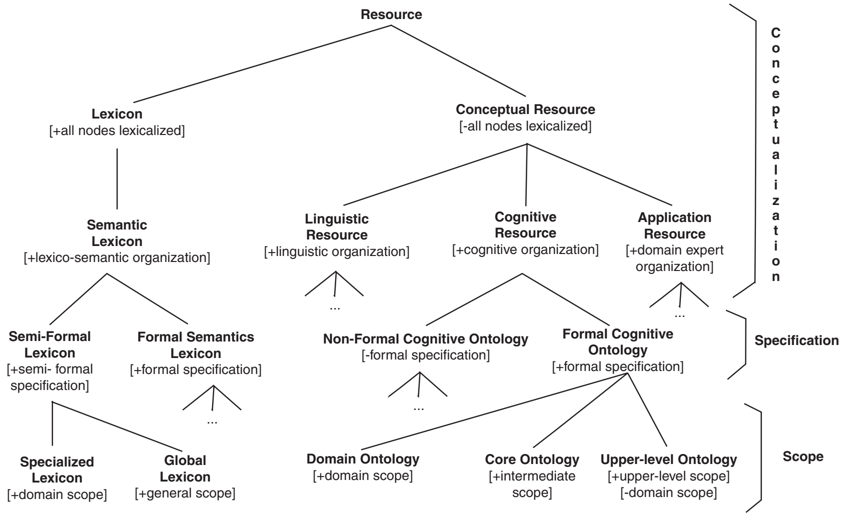

Figure 1.3 Ontolex resources taxonomy overview

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.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

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.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

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.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

to derive new meanings. It is important to bear in mind that neither regular polysemy (Copestake and Briscoe, 1995; Apresjan, 2000) nor creative use (Pustejovsky, 1995) can be exhaustively listed. The notion of word sense as a discrete semantic unit is itself put in question (Kilgariff, 1997). In the context of this book, we avoid this thorny issue with a data-driven approach. In the development of NLP and ontologies whether word senses really exist, or not, is not essential; but the frequent references to word senses in major existing resources makes them important elements to be considered. However, given the bottom-up approach, one still needs to deal with the different granularity among various resources. An interesting proposal for answering this issue is proposed in Chapter 15, where a method is proposed to have some control on the level of granularity of the sense introduced in the final resource.

1.2.2 Relations

In ontologies, concepts are integrated into a coherent whole with relations. The nature and the number of relations have been the subject of many studies in the field. In ontology, relations are conceptually driven and take concepts as arguments. On the other hand, lexical resources are concerned with the organization of lexicalized items. The relations they feature have only an indirect conceptual nature such as **antonymy**, which is primarily a relation between word *forms* and not between concepts or word meanings (Fellbaum, 1998: 49).[8] Relations with the same name in formal and linguistic ontologies might appear to be quite different under closer scrutiny. Moreover, the research issues involving these relations are quite different from the formal-ontology and lexical-resource perspectives. For formal ontologists it is important to clarify the nature and the formal properties of the relations: to which kind of entity do they relate (classes or individuals), are they reflexive, symmetric, transitive, and so on? For example, formal ontologists have focused attention on the **is-a** relationship overload. This relation has been used extensively but was often only loosely defined and merely corresponded to the intuitions triggered by its natural language expression **is-a**. On the other hand, relations of lexical resources hold between word senses (e.g. **hypernymy**) or even simply words in the case of lexical relations like **antonymy**. For lexical resources, the focus in recent studies is not on precise definitions for these relations, but more on the methods for discovering them automatically and for their application in extracting lexical knowledge. It is important to note that a general classification of these relations as either paradigmatic or syntagmatic is common to both conceptual and lexical approaches.

[8] In this case it is the conceptual opposition that can be associated with the lexical **antonymy**.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

1.2.2.1 Paradigmatic relations Paradigmatic relations hold between elements of the same nature that belong to a common paradigm. We restrict ourselves to terms that can be replaced in a given context as in Example 1.1. They belong typically to the same syntactic category as opposed to items related through syntagmatic relations.

(1.1) a. A(n) animal/dog/cat/dalmatian crossed the road.
b. He ate/devoured the small rabbit.

In Cruse, 1986, paradigmatic relations are associated with congruence relations between classes such as **identity**, **inclusion**, **overlap**, and **disjunction**. Indeed the best-known paradigmatic relations in the lexical domain are **synonymy**, **antonymy**, **meronymy**, **hypernymy**, and **hyponymy**. In ontologies, related conceptual relations of **conceptual opposition**, **part-of**, **is-a-kind-of** are formally defined. However, more relations can be thought of, for example the relations among the siblings in a taxonomy are sometimes good candidates (e.g. red/black/blue, or cat/dog). Such richness in paradigmatic relations (Murphy, 2003) leads to the proposal of a very general *principle of relation by contrast* that covers paradigmatic relations. Huang *et al.*, 2007 proposes a specific relation called **paranymy** to cover paradigmatic relations among concepts belonging to the semantic classification.

Until recently, these relations have been the ones most widely studied and applied. Many NLP uses of ontologies restrict themselves to the use of a simple taxonomy. Relations have received different names according to the framework considered. This terminological profusion suggests different concerns and perspectives. For example, the highly debated **hypernymy** (and **hyponymy** its inverse) relates lexical entities but has been often used as a straightforward relation between concepts. The relation between concepts is also called **is-a** relation although **is-a-kind-of** is less ambiguous and favoured by ontologists who equate it with **subsumption**.

1.2.2.2 Syntagmatic relations As mentioned above, *syntagmatic relations* hold between entities of different natures; the items related by these relations co-occur frequently but cannot be replaced by one another. They are often lexicalized by words having different syntactic categories. In lexical semantics, syntagmatic relations are more related to studies of syntax/semantic interface focusing on predication and thematic roles. Syntagmatic relations include relations between endurants (objects, agents) and perdurants[9] (including events and processes) at the lexical level (noun/verb relation), or, for example, between a category and its attributes (noun/adjective relation

[9] The term *occurent* is also used.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

at the lexical level). Many of these relations have linguistic counterparts as *case relations*. They have been studied in depth by Fillmore to develop his Case Grammar (Fillmore, 1968) and they constitute the majority of FrameNet (Baker *et al.*, 1998) relations (see also Chapter 4).

While WordNet and most recent ontological-based works have focused on paradigmatic relations,[10] syntagmatic relations received less attention for resource developers while their importance for NLP applications gradually appeared as essential.

Even though its development is quite recent, FrameNet, as well as the related theory of Construction Grammar, is now subject to the same attention given to WordNet by computational linguists and ontology builders. This complementarity between syntagmatic and paradigmatic relations in WordNet and FrameNet and their efficient combination is an important element of this applied research area.

The syntagmatic/paradigmatic distinction partially overlaps with the division proposed in Nirenburg and Raskin, 2001 between syntax-driven and ontology-driven lexical semantics. The former corresponds to syntagmatic relations whose study had been bound tightly to the syntax/semantics interface (Levin, 1993). The latter, although putting a strong emphasis on paradigmatic relations (and in particular taxonomies and meronymies), includes also relations belonging to the syntagmatic class (e.g. **participation** of objects into processes).

1.3 Theoretical framework for the ontologies/lexicons interface

The fields involved in knowledge representation, regardless of whether they have a declared objective of psychological adequacy or not, already have a rich heritage. Several approaches can be broadly distinguished: philosophical studies tracing back to Aristotle, psychological studies focusing on the mental representation of knowledge, and linguistic studies. The topic of the interface between ontologies and lexical resources is therefore a re-examination of traditional issues of psycho-linguistics, linguistics, artificial intelligence, and philosophy in the light of recent advances in these disciplines and in response to a renewed interest in this topic due to its relevance for the Semantic Web major applications.

[10] To be fair, WordNet does host some syntagmatic relations such as **cause** from the beginning. However, the coverage of these syntagmatic relations is not comparable with the extensive hierarchical network in WordNet made up of paradigmatic relations (see Chapter 2 for a quantitative presentation of the relation in WordNet.). Moreover, the initial design of WordNet with a distinct structure for each syntactic category precluded the development of cross-category relations such as the ones present in EuroWordNet (Vossen, 1998).

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

Overall, the importance of a multidisciplinary approach is recognized for lexical resources development[11] and knowledge representation as acknowledged by many influential contributions to the field (Hobbs *et al.*, 1987; Sowa, 2000; Pustejovsky, 1995; Nirenburg and Raskin, 2004; Guarino, 1998b).

This section explains how such a rich ground is an opportunity for the current research in NLP, knowledge representation, and lexical semantics. We are particularly interested in the interface between formal ontology and lexical resources or linguistic ontologies. Here formal ontologies are understood as ontologies motivated by philosophical considerations, knowledge representation efficiency, and re-usability principles. Lexical resources or linguistic ontologies have structure motivated by natural language and more particularly the lexicon.

1.3.1 The cognitive ground

1.3.1.1 Categorization Studies on categorization received a lot of attention from the cognitive side. The componential semantics (Katz and Fodor, 1963) in which the category of a word is defined as a set of features (syntactic and semantic) that distinguishes this word from other words in the language is one of the most influential accounts available. It is striking that this model fits extremely well with Formal Concept Analysis (FCA). Developed by Ganter and Wille (1997) and first applied to lexical data in Priss, 1998 and 2005, this framework is nowadays in use in several ontological approaches (illustrated, for example, in Chapter 6 of this volume). However, componential semantics has been limited by various developments centred on the notion of prototypicality (Rosch, 1973, 1978). It has been empirically established that the association of words and even concepts with a category is a scalar notion (*graded membership*). The problem of category boundaries, of their existence and of their potential determination, is therefore a serious one. Contextual factors have been said to influence these boundaries. Another issue is the use of a feature list that has been said to be far too simplistic and that raises the question of the definition of the features themselves. However, to see a parallel in the definition of categories from philosophy see section 1.3.2.

Besides the issue of prototypicality, another common ground is the investigation of the models for concept types (sortal, relational, individual, and functional concepts), category structure, and their respective relationships to 'frames'. There is wide converging evidence that language has a great impact on categorization. When there is a word labelling a concept in a certain language it makes the learning of the corresponding category by children much faster and easier.

[11] WordNet was originally intended for a psycho-linguistic experiment.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

There is much more to say about categorization, but we point the reader to Wierzbicka, 1990, Croft and Cruse, 2004: 77–92, Murphy, 2003, and Schalley and Zaefferer, 2006, in which these approaches and their limitations are discussed at length.

1.3.1.2 Predication In linguistics, a large body of work is focused on predication since it directs sentence interpretation. These works have been pioneered by Fillmore (1976) who proposed that we should analyse words in relation to each other according to the frame or script in which they appear. The study focuses on relations expressed by grammar case (Fillmore, 1968). In this domain essential contributions on argument structure (Grimshaw, 1990), thematic roles, selectional restrictions (Dowty, 1990), and type coercion (Pustejovsky, 1995) have been made in recent years. This field of research produced resources such as FrameNet (Baker *et al.*, 1998) and VerbNet (Kipper *et al.*, 2000).

1.3.1.3 Conceptual and semantic level: Are they identical? Many proponents of the cognitive approach to languages postulate that semantics is equated with the conceptual level. Jackendoff explains that surface structures are directly related to the concepts they express by a set of rules. These concepts include more information associated with world knowledge (or encyclopedic knowledge). Since, according to him, it is impossible to disentangle purely semantic from encyclopedic information without losing generalizations (Jackendoff, 1983), semantic and conceptual levels must form a single level.

However, Levinson (Gumperz and Levinson, 1996; Levinson, 1997, 2003) advanced serious arguments, involving in particular pragmatics, in favour of the distinction between semantic and conceptual representations. These differences are explained by different views on language inherited from different theoretical perspectives. While Jackendoff focuses on Chomsky's I(nternal)-language, Levinson insists on the importance of the social nature of language and therefore takes care of the rather neglected E(xternal)-language in Jackendoff's account. Language as primarily a tool for communicating rather than a tool for representing knowledge (in someone's mind) corresponds to these different perspectives.

From an applicative viewpoint, Bateman (1997) argues, on methodological grounds, for the need of an interface ontology between the conceptual and surface levels. He specifies that such a resource should neither be too close to nor too far from the linguistic surface and details the General-Upper-Model (Bateman *et al.*, 1995) as an example of such a balanced interface ontology. This is also the line followed by Nirenburg and Raskin (2004) and exemplified in Chapter 7 of this volume. Pustejovsky and his colleagues, on the other hand,

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

follow the direction of a single structure, though highly dynamic, as in the generative lexicon (Pustejovsky, 1991, 1995).

1.3.2 The philosophical ground

Determining a system of ontological categories in order to provide a view as to *What kinds of things exist?* is one of the essential tasks of metaphysics. It is also probably philosophy that offers the highest number of (sometimes contradictory) propositions regarding this issue, from Aristotle to contemporary authors. These proposals can differ strongly both on the nature of the ontological categories (*How exactly are the categories delineated?*) and on actual ontological decisions (e.g *What are artefacts, organizations, holes...?*). In this context the focus has been mainly on the upper level of the knowledge architecture and in finding an exhaustive system of mutually exclusive categories (Thomasson, 2004). The lack of consensual answers on these matters has resulted in a certain scepticism with regard to the determination of such an ontological system. However, recent approaches aiming at taking the best of philosophy while avoiding its pitfalls are rendering philosophical considerations worth the exploration.

The crucial aspect where philosophy can help the ontology builder might not be the position of such and such a concept in a ‘perfect’ taxonomy, but rather the methodology for making such decisions. The second important aspect is grounded on Strawson’s distinction between *revisionary* and *descriptive* metaphysics. For Strawson (adopting a descriptivist stance), the material accessible to the philosopher is not the real world but how the philosopher perceives and conceptualizes it. Contrarily, a revisionary approach uses rigorous paraphrases for supposedly imperfect common-sense notions. See, for example, a discussion about the difficulties that such a revisionary approach meets when trying to deal with objects as simple as holes (Casati and Varzi, 1996). Revisionary approaches tend to discard natural language and common-sense intuitions as serious sources for ontological analysis. On the other hand, the descriptivist stance is presented to be safer philosophically. It also provides a solid methodological starting point for practical ontological research. More precisely, by allowing different ontologies (as different descriptions of the world) to co-exist, it is able to avoid never-ending ontological considerations on the real nature of the objects of the world.

This move leads to the distinction between Ontology as a philosophical discipline and ontologies as knowledge artefacts we are trying to develop (Guarino, 1998b). Modern ontology designers are not looking for a perfect ontology but consider many potential ontologies concerning different domains and capturing different views of the same reality.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

To succeed in this task, philosophical works try to determine a set of fundamental categories. In Thomasson, 2004 three main methods have been summarized:

- *The feature negation method*: This method ensures exhaustiveness and mutual exclusivity. This method can divide any category into two others by distinguishing the things having a certain feature from those not having it. The feature system can be developed in a multi-dimensional fashion forming a matrix (potentially very rich) more than a tree. The two main philosophical issues with this approach are (a) the negative characterization of many concepts that might end up in the same category without sharing any intrinsic property and (b) the inadequacy of many feature/category pairs established by linguistic tests.
- *The absurdity detection method*: This method, originating with Husserl and Ryle (Ryle, 1971), consists of testing an expression (grammatical context) in which a concept can be employed without exhibiting an absurdity. For example, *this house is brown* is correct while *this hypothesis is brown* is absurd, thus exhibiting the different nature of these two concepts. This method can be used to prove that two elements belong to the two categories, but cannot prove that two candidates belong to the same category, since a new dividing context may always be found.
- *The method of distinguishing Identity and Existence conditions*: This method is a variation and an extension of the previous one. Instead of considering expressions, it is the objects denoted that are examined. According to Frege (1884), *names* are different from other terms by their condition of identity added to their criteria of application (that corresponds to the previous method). Names are associated with *sorts* of thing (*sortal term*). Armed with these names, it is possible to distinguish different categories of *objects* as correlates of names, thus distinguishing ontological categories. Then, the application conditions and identity criterion are tested on the sortals. The second essential ingredient for identifying the categories are the identity conditions. On this issue it is Quine's recommendation 'No entity without identity' (Quine, 1960) that forces us to know how to identify (and therefore distinguish) a given entity from another one before bringing it to existence. For example, an ANIMAL is distinguished from the COLLECTION-OF-PARTICLES constituting it because they each have a different *Identity Condition* (IC). All the particles are involved in the identity of the collection but not of the animal for which the IC has something to do with its DNA. Loosing a leg or a hair does not affect the identity of an animal but it does affect the identity of the corresponding collection of particles (which becomes a new collection of particles).

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

The *OntoClean* methodology (Guarino and Welty, 2000a, 2002a), widely accepted and now commonly applied to various practical situations, draws upon the last method and proposes a powerful methodology for building and maintaining ontologies. From the traditional philosophical investigations, *OntoClean* proposed a set of useful meta-properties that apply to the categories (called properties); these meta-properties include the identity condition (Guarino and Welty, 2002b). As emphasized earlier, the objective of this methodology is not to propose *the* only set of acceptable categories, but to help ontology builders and maintainers to make explicit their ontological commitments. This helps to discover ontological issues early in the building process rather than suffering from it once the ontology is developed. Said differently, *OntoClean* does not tell the designer which are the hypothetical 'perfect categories', but helps to make explicit the consequences of the choices made during conceptualization and specification. Chapters 3 and 10 of this book present some details of this approach.

1.3.3 The lexicographic ground

Lexicography, the discipline dealing with dictionaries, is traditionally split into the practical aspects, the craft of creating dictionaries, and the more theoretical lexicology. However, serious lexicography works generally combine practical and theoretical aspects. Since traditional dictionaries are the ancestors of our computational lexical resources, it might not be so surprising that early lexicography studies have already met with most of the difficulties that modern approaches are facing nowadays. Such early studies included the characterization of the different types of dictionaries (Shcherba, 1995), which appears quite in line with the discussions of the previous sections as illustrated in the following list of essential distinctions that operate among dictionaries:

- Encyclopedic vs general dictionaries. Encyclopedias emphasize the importance of proper nouns and show that many of them must be included in general dictionaries. Under this distinction, Shcherba also discusses the words that receive quite different entries in dictionaries and encyclopedias because of their respective level of specificity.
- Form-based (*ordinary*) from meaning-based (*ideological*) dictionaries. In the latter, the importance of synonymy and other relations was emphasized and exemplified through Roget's thesaurus.
- *Defining* from *translating* dictionaries where the complex nature of the relation between lexical entries across languages is emphasized.

Practical lexicography is restricted by the fact that lexica are essential resources which in turn require deployment of significant resources to develop.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

For instance, timely development of a dictionary often requires deliberation of economy and exclusion of non-essential entries. However, far from being simply a burden, these constraints force the lexicographer to meaningful choices in order to include only lexical information corresponding to the general language and rejecting the rest into encyclopedias and specialized lexicons.

The importance of the distinction between lexical and encyclopedic knowledge is also fundamental for Nunberg and Zaenen, 1992. They explain that when language is taken as a social artefact in which cultural context is an active element of its definition, many regularities from encyclopedic knowledge deserve to be integrated in dictionaries.

Despite its practical orientation, lexicography can greatly benefit from linguistic theory as advocated in Apresjan, 2000, where four trends of modern linguistic theory are taken to be of immediate relevance for ‘systematic lexicography’: (a) the search for a common-sense picture of the world, (b) the shift from study of separate words to larger lexicographic types, (c) the meticulous study of separated word senses in all of their linguistically relevant aspects, (d) the convergence of grammatical and lexicological studies towards an integrated theory of linguistic description. All these elements also constitute the ground for the generative lexicon theory (Pustejovsky, 1991, 1995) that is presented in the next section.

1.3.4 The linguistic ground and contemporary lexical semantics

The frameworks presented in the previous sections largely included linguistic considerations and contributions from linguistics. However, the goal was generally not the study of language or the development of natural language systems. With regard to the philosophical ground, natural language semanticists and philosophers of language have been deeply interested in natural language ontology, or to which ontological categories the study of natural language commits. The work on tense and aspect (Dowty, 1977) but also on plurals (Bach, 1986a; Link, 1983) or mass/countable nouns provides research materials for a field that has sometimes been called natural language metaphysics (Bach, 1986b; Dölling, 1993). Although such crucial studies formed a firm theoretical ground allowing future advances, this body of research had little direct impact on the development of practical resources.

An important exception to this is the generative lexicon of James Pustejovsky (1991, 1995). The generative lexicon is a ground-breaking contribution to the study of lexicon and its relation with ontology. It combines the philosophical ground with a rich lexical-semantic theory featuring a sophisticated account for predication. The generative lexicon addresses two serious issues

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

encountered by traditional resources. First, they are generally based on sense-enumeration and attempt to include a list of word senses. Pustejovsky argued convincingly that word senses are infinite since language producers can easily create new senses for a word. For example, coercion operated by the predicate alters the semantic type of a given argument. In this context the lexical entries are more dynamic and correspond to combinations of several aspects of meaning, the *qualia roles* inspired from Aristotle's *mode of explanations*. For nouns, the different aspects of meaning are *constitutive* (what are the parts of this object), *formal* (what is this object, the classical taxonomy), *agentive* (what is its origin) and *telic* (what is its intended function). The multidimensional view of meaning is sometimes captured by allowing for massive multi-inheritance in the hierarchy. However, such practice is considered precarious from a knowledge representation viewpoint since the structure becomes murky, and it is more difficult to preserve the consistency of the whole. The generative lexicon resolves this dilemma and sanctions a multidimensional view of meaning by maintaining a sound knowledge organization through the use of different relations rather than a single entangled tree. For example, inheritance is not allowed along all the dimensions and only careful orthogonal inheritance is allowed (Pustejovsky and Boguraev, 1993).

The generative lexicon has received a lot of attention, and some practical resources implement its principles into real-scale lexicons such as SIMPLE (Lenci *et al.*, 2000) including an extended qualia structure (Busa *et al.*, 1999) that is necessary to characterize precisely how each quale contributes to the typing of a concept. This powerful research on predication and coercion so far suggests that the four qualia aspects do not exhaust all the aspects of meaning coercion. For example, Asher and Denis (2004) proposed that we should generalize the idea of the generative lexicon to an arbitrary set (and not only the four qualia) of relations contextually triggered (including discourse information).

This section closes this brief presentation of the theoretical background supporting this volume. Before presenting in detail the structure of the book, the next section emphasizes the bidirectional nature of our enterprise and the centrality of NLP.

1.4 From ontologies to the lexicon and back

An essential view defended in this book is the bidirectionality of the relation between ontologies and lexical resources. We reject the primacy of ontological research over lexical research or vice versa. We argue for a balanced combination following clear principles backed by theoretical investigations. We consider both how ontological methodology and knowledge can enhance lexical resources and how these resources can in turn benefit ontologies.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

1.4.1 From ontologies to lexicons

This direction concerns the ontological enhancement of lexical resources by using them as consistent knowledge sources for knowledge engineering applications. It includes checking the knowledge structure of lexical resources and suggests improvements of these resources (see Chapter 3). This direction is also concerned with the axiomatization of the glosses found in lexical resources (Harabagiu *et al.*, 1999; Gangemi *et al.*, 2003a). Some contributions presented here (see Chapters 3) argue for performing this integration as a preliminary step and only in a second step inputting lexical data in the ontology. However, other proposals apply more straightforwardly formal tools such as Formal Concept Analysis on natural language resources (see Chapter 6) or use directly conventionalized linguistic objects such as the Chinese writing system as potential linguistic ontologies (see Chapter 8).

1.4.2 From lexicons to ontologies

The other direction (from lexical resources to ontological knowledge) has been so far primarily concerned with the population by lexical terms of ontological resources. This process is an efficient way of building vast ontologies of common-sense or domain knowledge (according to the vocabulary included). Several experiments of this kind are represented in this book, either for populating an existing upper level (see Chapter 2) or for creating complete domain ontologies (Chapter 17). The existing lexical resources might also be used as guidelines during the construction of the ontology.

A general result of the contributions presented in this book is the improvement of the two types of resources by providing systematic links between them. Several models are compared in Chapter 10, and some promising solutions are presented in Chapters 12 and 13.

1.4.3 The centrality of natural language processing

According to the advocated bidirectionality, NLP is seen both as an application and as a tool for the ontolex interface (as emphasized in Chapters 7, 10, and 16). Several chapters illustrate the use of NLP techniques for building ontologies (semi-automatically). The techniques used range from syntactic parsing and semantic analysis to term and relation extraction. Chapter 6 points towards a deeper use of syntactic and semantic analysis for improving the ontology acquisition results. But NLP also makes use of the ontologically enhanced lexical resources for a variety of classical NLP tasks. Such applications include information retrieval and question-answering (see Chapter 15 and 16), co-reference resolution, and more globally semantic analysis for deep-language understanding (Chapter 7).

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

1.5 Outline of chapters

The book is divided into four parts. The first part is composed of five chapters. Chapters 2, 3 and 4 introduce essential resources of our field and expose different approaches of the combination between formal ontologies and WordNet or FrameNet. The part is concluded by a roadmap for the ontolex interface that provides perspectives for future work in the field (Chapter 5).

The second part, also composed of five chapters, concerns the discovery and representation of conceptual systems. This part of the book shows the variety of methods that can be used for unveiling ontological knowledge as well as the variety of representations. The sources for unveiling knowledge might be surprisingly rich as exhibited in this section, from automatic learning techniques coupled to formal representation tools (such as Formal Conceptual Analysis in Chapter 6) to the investigation of a 3,000-year old body of knowledge such as the Chinese writing systems described in Chapter 8. This chapter investigates further the notions of conceptualization and specification delineated in the general introduction in order to explain the different ways taken first for establishing a shared understanding of a domain (conceptualization) and then for specifying it in a given language (formal or not). The representation of a knowledge system is also addressed with the firm position of Chapter 7 defending a structure for semantic, ontological, and factual information. This chapter also shows how such architecture can be used in a practical treatment of lexical and compositional semantics of events of change. Finally, Chapter 9 proposes an ontological framework for cognitive linguistics.

The third part of the book, composed of six chapters, addresses explicitly the theoretical and practical problems encountered when interfacing ontologies and lexical resources. Chapter 10 provides a methodology classification that is also used for positioning the remaining chapters of this part. The other chapters include the presentation of SINICA-BOW (Chapter 11) an English–Chinese bilingual resource that combine a lexical resource (WordNet) and an ontology (SUMO). Chapter 12 discusses, on the one hand, the traditional model of semantic lexicons in which senses are assigned to lexical items and the set of senses is mostly open-ended. On the other hand, ontologies are said to provide formal class definitions for sets of objects, which can be seen as a ‘sense’ for those lexical items that express such objects. The model described in this chapter aims at merging these two disparate views into a unified approach to lexical semantics and ontology-based knowledge representation. Then Chapter 13 concerns the combination of linguistic ontologies having different granularity. This work tries to go a step further in the direction of the interoperability of specialized linguistic ontologies, by addressing the problem of their integration with global ontologies. This scenario offers some simplifications with respect to the general problem of merging ontologies, since it enables us to define

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press

a strong precedence criterion so that terminological information overshadows generic information in the case of conflicts.

The last part of the book concerns more explicitly NLP issues. The contributions (eight chapters) show the bidirectional nature of NLP at the interface between ontologies and lexical resources. NLP is used both for learning ontological knowledge and uses this knowledge in applications such as question answering, information retrieval or anaphora resolution.

Chapter 15 presents the Omega ontology, a shallow, lexically oriented term taxonomy. The chapter explains how such a resource is useful in a wide variety of applications such as question answering and information integration. Chapter 16 is devoted to the acquisition of lexico-semantic knowledge for question answering and evaluates the benefits of this acquisition. Chapter 17 is concerned with the semi-automatic construction or improvement of existing resources based on text or traditional resources.

https://doi.org/10.1017/CBO9780511676536.002 Published online by Cambridge University Press