Chapter 5
Ontologies and Multilingualism

Gilles Falquet and Jacques Guyot

5.1 Introduction: Ontologies and Natural Languages

The definition of an ontology as a specification of a conceptualization of a domain is independent of the terminology used in a particular natural language to describe this domain. In fact we can make a clear distinction between the conceptual structure of a domain and the way the concepts are designated by terms in a natural language. This view is exemplified in ontology specification languages such as OWL in which there is no connection with terms or texts in natural language, except for comments. In such a language, an ontology designer can arbitrarily define new concepts that do not correspond to any term in an existing language.

So why do we need to consider natural languages when building ontologies? There are multiple answers to this question, some of which are highly practical while others have a more theoretical background.

5.1.1 Theoretical Connections

On the theoretical side one can first observe that the lexicon of each natural language provides a conceptualization of the world. Most of the lexical forms, in particular nouns, designate a family of individuals that form a concept (e.g. dog, road, computer, …).

G. Falquet (✉) • J. Guyot
Centre universitaire d'informatique, University of Geneva, Route de Drize 7,
CH-1227 Carouge, Switzerland
e-mail: Gilles.Falquet@cui.unige.ch; guyot@cui.unige.ch

G. Falquet et al., *Ontologies in Urban Development Projects*, Advanced Information and Knowledge Processing 1, DOI 10.1007/978-0-85729-724-2_5,

This designation can of course be ambiguous in presence of polysemous forms like *bank* or *table*. The world's conceptualization generated by a language's lexicon is usually represented in lexical ontologies like WordNet, that are often used as a basis or skeleton for building more specific or formal ontologies. They are also of great help for many practical applications like synonym removal, word sense disambiguation, query expansion in information retrieval, etc.

Another theoretical connection between ontologies and natural languages originates in the non-circularity of definitions. It is usually desirable to avoid circular definitions in formal ontologies. But the only way to avoid circularity is to admit that some concepts, called primitive or basic, are not defined within the ontology. Then, the only way to know what these concepts are is either to name them according to a well-known natural language term or to describe them with words. For instance, the CityGML model, in its *Water Bodies* sub-model refers to water body classes such as lake, river, ditch, bayou, etc., that are not defined in the model. This is acceptable because the purpose of this ontology is to describe urban objects and these descriptions do not require extremely precise definitions for concepts that are on the border of the domain. In this case, the linguistic form, like *sea*, is associated to a consensual meaning that is considered as sufficient.

Finally, linguistic forms are the only way to anchor an ontology in a real domain. An ontology whose concepts and relations identifiers are purely arbitrary strings of characters (C419, C2001, icl, pof, …) would hardly be considered a conceptualization of some domain. At some point there must be a link between the "internal" concept identifiers and some known concept of the domain. This is where linguistic forms play an important role.

5.1.2 Practical Connections

Ontology designers must base their work on solid foundations, usually provided by domain specific information sources such as dictionaries, reference texts, legal texts, and many other types of documents. These documents, except for pictures, are expressed in some natural language. Moreover, in every specialized domain of human activity, a specific terminology has emerged to easily and unambiguously designate the frequently used concepts. Because specialists of the domain have learned to work with these concepts, it is quite clear that any usable ontology should be consistent with this terminology and the conceptualization it induces.

Similarly, from the ontology designer point of view it is certainly more convenient to work with concept names that exist in the natural language, even if the concept meaning in the ontology differs from its usual sense in everyday language. At some point the designer may also be led to create new concepts, acting as a terminologist, here again it is often suitable to name these concepts with (combinations of) existing linguistic forms.

5.1.3 Multilingualism

When working in a multilingual environment, the above-mentioned connections between an ontology and a natural language must be extended to several natural languages (Collier et al. 2006). This may occurs is several circumstances, for instance

- An ontology may serve as a common reference for an international community of users. In such a situation users generally prefer to accessing the ontology in their own language; they also need to find equivalent terms in other languages, e.g. for translation purpose.
- In ontology driven user interfaces, such as guided interactive information retrieval systems, the user will certainly be more efficient in her own language.
- In semantic indexing of large multilingual text corpuses (see Sect. 5.3 below) it is necessary to know the lexical form corresponding to a concept in all the considered languages
- The information sources required to build an ontology may exist only in some languages therefore the development process must take into account several languages (to avoid the reductionist approach consisting in translating all into a single target language)
- When an ontology needs to be localized, i.e. adapted to a particular language and culture, the ontological work should be carried out in several languages

Each one of these situations poses challenges of which we will explore some in the remaining of this chapter. We will first study the representation issues (how to take into account multiple languages when building ontologies), then, we will show how ontologies, connected to multilingual lexicons, can enhance information indexing and retrieval in a multilingual context.

5.1.4 Ontologies and Point of Views

In a context where different point of views must be taken into account, it can be useful to consider each point of view as a different language. For instance, it is well known that domain specialists have developed specific vocabularies to exchange information in a precise and non-ambiguous way. As a consequence, when a human activity spans several domains, the involved actors may experience communication problems due to this diversity of vocabularies. This can typically occur in urbanism related activities, such as urban planning, where urban engineers, architects, politicians, transportation engineers, or citizen organizations participate in decision processes. Since each one of these groups possesses its own vocabulary and conceptualization of the world, improving communication between them cannot rely on the development of a single "monolingual" ontology. In fact, we are confronted with a situation that is similar to multilingualism or multiculturalism. In particular, the "near synonym" problem frequently arises as well as differences in definitions of the same concept.

5.2 Approaches to Multilingualism in Ontologies

5.2.1 The Basic Concept-Centric Approach

This approach is based on the idea that most of the domain concepts exists in all the considered cultures. In other words, concepts are universal while their linguistic representation is culture-specific. Admitting this hypothesis, multilingualism can be supported by first building a "universal" ontology and then associating linguistic information to each concept.

The OWL ontology language proposes a basic mechanism to handle linguistic information in the form of *annotation properties*. An annotation property is a kind of meta-data attached to a concept. Its value is a string together with a language tag. In OWL knowledge bases the rdfs:label property is typically used to provide the linguistic form of a concept in different languages. Figure 5.1 shows the forms for the concept *Piéton* in French, English, and Italian (in the Protégé ontology editor).

Many existing ontologies are based on this approach. For instance, the Unified Medical Language System (UMLS) (National Library of Medicine 2009) is comprised of a set of concept identifiers (over one million) associated to terms originating from sources vocabularies from 18 different languages.

The concept-centric approach is well suited for normative terminologies, e.g. for ensuring that the same term is always translated in the same way in all the official documents issued by an organization. In a sense, these ontologies are similar to multilingual thesauri, the aim of which is mostly to define a controlled vocabulary. The main disadvantages of this approach are

1. The lexical information attached to a concept is limited to a character string, so there is no possibility to define relationships between lexical forms or to build sophisticated lexical structures.
2. The lexical forms (labels) are strictly equivalent, i.e. each label of a concept is supposed to designate exactly this concept. This can be true for very specialized domains but that is rarely the case for wider domains. For instance, the usual translation of the French word *fauteuil* (armchair) into Italian is *poltrona* but their meanings are slightly different (a *poltrona* is necessarily perceived as comfortable which is not the case for *fauteuil*). If it is necessary to be really precise

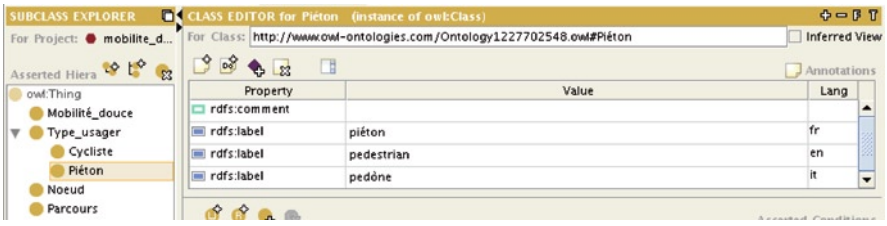

Fig. 5.1 Lexical forms attached to the Pieton concept in three different languages (in the protégé ontology editor)

then one must create two different concepts and use (invent) new terms to designate them in the language in which there is no direct lexicalization for them.

5.2.2 Concept-Centric with Structured Linguistic Elements

A more sophisticated version of this concept-centric approach can be obtained by considering à three-level model where concepts, terms, and forms are represented.

The conceptual level is intended to represent the concepts (or meanings) and their definitions. It is comprised of ontological elements such as concepts, semantic relations, properties, individuals. Formulae or texts express the concept definitions and domain axioms.

The terminological level is made of terms, which are associations between concepts and lexical forms. For instance the chemical term *acid* associates the linguistic form acid to the concept defined as *a compound which donates a hydrogen ion to another compound in a reaction*. Terms may be interrelated through terminological relations such as antonymy.

The lexical level represents the forms, which are character strings used in written language. These forms may be connected through lexical relationships such as plural or other inflectional variants. Moreover, additional relations and categories may be defined: variants, notes, context, etc.

There is, for instance, a proposal to re-implement the AGROVOC multilingual thesaurus in OWL with such a structure (Lauser et al. 2002, 2006; Soergel et al. 2006). In this case the ontology has two main concepts: *domain_concept* and *lexicalization*. All the domain concepts are subconcepts of *domain_concept*, while terms are instances of *lexicalization*, and forms are (string) properties of terms. Terms may have properties like *has synonym* or *has translation* that link them to other terms.

The multilingual support proposed in the Neon project (Montiel-Ponsoda et al. 2008) extend this approach by proposing a sophisticated structure to represent lexical information and to link this information to ontological element of the OWL language (class, property, individual, …). The aim of this model is to fully localize an ontology, so that an ontology engineer or a user can work in his or her language. This is why every ontological element must have a localized lexical form.

The sophistication of the terminological level remedies the problem of strict equivalence of terms that exists in the basic concept-centric approach. Indeed, it becomes possible to associate weights to the links between terms and concepts, to indicate preferred terms, etc.

5.2.3 Interconnection and Alignment Approach

Instead of considering a unique ontology that represents the domain conceptualization, it is possible to maintain individual ontologies, corresponding to multiple views of the domain, and establish equivalence or similarity links between their concepts.

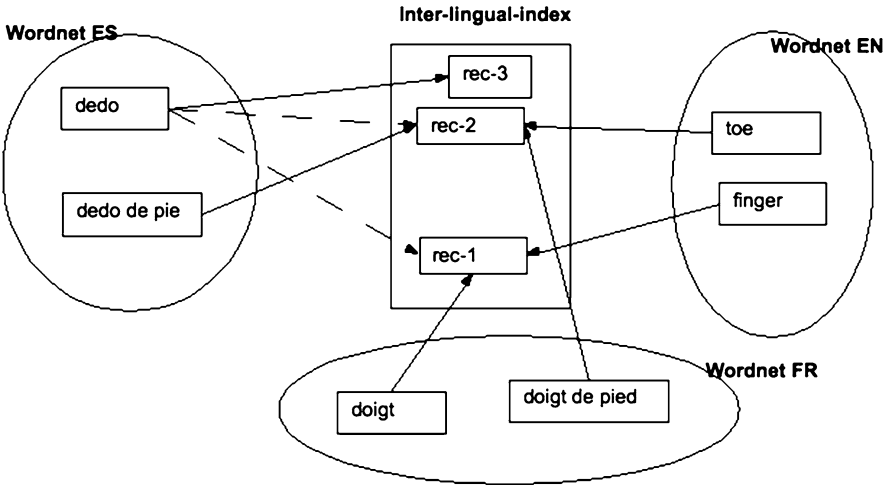

Fig. 5.2 Interconnection records between synsets of different wordnets. *Solid lines* represent EQ_SYNONYM relations, *dashed lines* represent EQ_HAS_HYPONYM relations

If the concepts correspond to terms in different natural languages, this is a mean to keep the different conceptualizations of the world and not to impose a unique view. This is particularly useful for lexical ontologies that are bound to their source language.

The creators of the EuroWordnet initiative have taken this approach to interconnect Wordnet ontologies developed for different languages. Their interconnection model is based on a so-called "inter-lingual index" (ILI). The ILI is a set of ILI records that are intended to connect equivalent concepts. All the concepts belonging to different ontologies that are linked to the same ILI record are considered as equivalent, as shown on Fig. 5.2. However, as mentioned for the previous approaches, the equivalence notion is often too restrictive. It often happens that a term in one language has no exact equivalent in another one. To address this issue the ILI has been extended in two ways:

1. The initial set of ILI records, which was directly drawn from the English Wordnet (i.e. there was a one to one correspondence between ILI records and English synsets) has been extended with new records that represent specific concepts of other languages. For instance, the Spanish word *dedo*, which means finger or toe, has no corresponding term in English. Thus a new ILI-record for dedo must be created.
2. Different kinds of relations between a synset and an ILI-record have been introduced (Peters et al. 1998):

   EQ_NEAR_SYNONYM when a sysnset matches multiple ILI-records.
   EQ_HAS_HYPONYM when a synset is more general than all available ILI-records.
   EQ_HAS_HYPERNYM when a synset can only be connected to more specific ILI-records.

This interconnection approach preserves the conceptual structure of each ontology. However, it requires a very precise and tedious work, carried out by terminologists, to establish the interlinking structure.

When the ontologies are more formal it becomes possible to automate the interconnection phase by applying concept similarity measures, see for instance Rodriguez and Egenhofer (2003) or ontology alignment techniques such as the one proposed by Li et al. (2006). These methods are based on structural comparisons of the concept definitions (how they are related to other concepts and where they are in the concept hierarchy) and on textual comparison of the comment, glosses, or terms associated to the concepts (with the help of multilingual dictionaries). They are appropriate for providing a first alignment of the ontologies, which must be followed by a human revision phase to improve the quality of the alignment.

5.3 Applications of Multilingual Ontologies

5.3.1 Finding and Checking Translations

When working in a very specialized domain, human translators and terminologists usually don't find term translations in existing multilingual dictionaries or thesauri. In addition, they must ensure that the terms they use really have the intended meaning. Multilingual ontologies made of aligned or partially aligned monolingual ontologies may be of great help in such situations.

For instance, Falquet and Mottaz (2000) propose a semi-automated technique to find the best candidate translations for a term. Given two monolingual ontologies A and B, the first phase consists in explicitly aligning the basic concepts of both ontologies, i.e. those concepts that are not explicitly defined in their ontology. Generally these basic concepts are not central in the domain and so deciding if two such concepts are equivalent or have subconcept relation is relatively straightforward. For instance, an urban ontology may refer to the concepts *color*, *air*, or *tree* without defining them explicitly. Figure 5.3 shows two concept definitions (for *armoire* in a

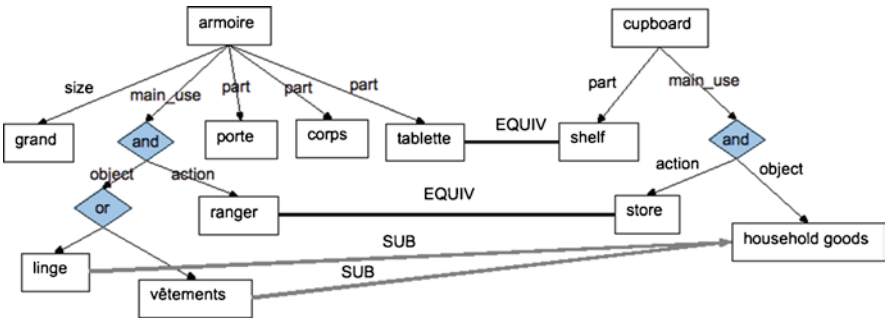

Fig. 5.3 Two concept definitions with aligned basic concepts

French ontology and *cupboard* in an English ontology) together with the aligned basic concepts they refer to. The second phase makes use of these basic equivalences to compare the definitions of defined concepts. It computes an edit distance between a definition *a* in A and a definition *b* in B by counting the number of change operations needed to transform *a* into a definition *a'* that is equivalent to *b*. The candidate translations for a concept are the concepts of the other ontology with the most similar (closest) definitions.

5.3.2 Multilingual Information Retrieval

Multilingual information retrieval (MLIR) consists in finding the most relevant document for a user need, considering that

1. the information need is expressed by a set of keywords or key phrases or sentences in the user's own languages
2. the document corpus contains documents written in different natural languages

MLRI has become more and more important with the advent of new communication technologies that enable users to access remote information sources. In many occurrences, these sources may contain documents that are not written in the user's preferred language but in some other language the user understands or for which he or she can afford a translation. MLRI is also crucial in international organizations that often have several working languages or that produce translated versions of their documents.

A classical approach for solving MLIR requests proceeds in three steps:

1. automatically translate the query into all the supported languages;
2. match each translated query to the documents written in the same language (applying standard monolingual IR techniques);
3. merge the result sets (ordered lists of documents) to produce a single ranked list of relevant documents.

This last phase is particularly difficult because merging ranked sets cannot be carried out by a simple comparison of the relevance values (Reference) since they have been computed on different sets of documents.

With a multilingual ontology it becomes possible to handle the MLIR problem differently. The basic idea is to replace each term that appears in a document or in the query by a concept identifier. Then it is possible to apply mono-lingual IR techniques, simply replacing the word space by the concept identifier space.

Depending on the degree of sophistication of the ontology different types of processing can be achieved. The strict minimum is a flat list of concepts identifiers, each one with its lexical form in each language, this is in fact a kind of multilingual lexicon. Experiments have shown that this can be sufficient to provide acceptable results (Guyot et al. 2006). In addition, it is much easier to find multilingual lexicons (lists of words together with their translation) than fully formalized multilingual ontologies.

It is however clear that a more sophisticated multilingual ontology, with a multiple lexicalisations for each concept should improve the quality of the indexing process.

If a multilingual ontology with semantic relations (in particular *subconcept* links) is available then the ontology can serve to enhance the retrieval process in several ways:

Disambiguation. Although experiments have shown that disambiguation is less crucial than can be thought at first, it is obvious that indexing an ambiguous form (e.g. *table*) with the correct concept is always suitable. There exist several disambiguation algorithms that are based on the inspection of related terms in the ontology. For instance, if the words *chair* and *eat* are found near *table* in the text, this will indicate that the correct sense for table is probably *a piece of furniture having a smooth flat top* …, because this sense is close (in terms of semantic path) to senses for *chair* and *eat* in the ontology.

Reasoning. The matching process may take advantage of semantic relations determine that documents that do not match the query at the keyword level are nevertheless relevant. For instance, if the query is the set of keyword Q = {*bird*, *car*}, a document containing the words *sparrow* and *limousine* should be considered as relevant because the corresponding concepts are subsumed by *bird* and *car*. Other semantic relations such as *is_part_of* may also be used to enhance the matching process, depending on the context.

Interactive search. Interactive search techniques, such as faceted search, propose to build the user query by navigating within (subsets of) the domain ontology. By following semantic links the user should be able to discover the concepts that best fit her information needs and then access the documents that are indexed by these concepts. Since the interface must display the linguistic forms that denote concepts, not internal concept identifiers, it is clear that these techniques work only with ontolgies that have a (multilingual) lexical layer.

5.3.3 Semantic Annotation of Documents

The next generation of search engine should rely on semantic web techniques such as semantic annotation of documents. A semantic annotation, in its simplest form, is a list of concepts belonging to a domain ontology. The concepts associated to a document indicate what the document is about. This is similar to the semantic indexing process describe here-above. In this case the syntactic structure of the sentences is lost. In fact, this approach considers documents as bags of concepts and cannot rely on deeper semantic information.

A more precise kind of annotation consists in semantic graphs, for instance RDF graphs. In this case the graph nodes correspond to individuals that are concept instances and the labeled edges represent semantic relations between these individuals. The graph is thus a (partial) representation of the semantics of the document.

Terminologically rich and multilingual ontologies play a key role to enable semantic annotation.

1. They serve as references for labeling the graph nodes (with concept identifiers) and the graph edges (with relation identifiers).
2. Automatically annotating large collections of documents requires natural language processing tools (in particular parsers) to recognize the lexical forms corresponding to concepts and concept instances. These tools must be provided with adequate lexical information.
3. Natural language processing tools can take advantage of ontological knowledge to solve syntax analysis problems. For instance, ambiguous sentences may be disambiguated if some domain knowledge is available.

5.4 Conclusion

There exist natural and unavoidable connections between ontologies and natural languages. With the exceptions of ontologies that are used in fully automated processes that do not communicate with human users and do not access textual data, most ontologies must supply terminological information. This is particularly true when they are intended for multilingual context of use. We have seen that there are three main approaches to equip ontologies with multilingual terminological information: from simple concept labels to sophisticated terminological/lexical structures or ontology alignment techniques.

Multilingual ontologies certainly have an important role in knowledge engineering, in particular for applications that must deal with formalized knowledge *and* knowledge expressed in natural languages. We have presented three such applications: translation checking, multilingual information retrieval and the semantic annotation of documents.

Bibliography

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R., Takeuchi, K., Kawtrakul, A.: A multilingual ontology for infectious disease outbreak surveillance: rationale, design and challenges. Lang. Resour. Eval. **40**, 405–413 (2006)

Falquet, G., Mottaz Jiang, C.L.: Conflict resolution in the collaborative design of terminological knowledge Bases. In Proc. EKAW 2000 (International Conference on Knowledge Engineering and Knowledge Management), Lecture Notes in Computer Science.. Springer-Verlag, Berlin (2000)

Guyot, J., Radhouani, S., Falquet, G.: Conceptual indexing for multilingual information retrieval. In: Peters, C., et al. (eds.) Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Lecture Notes in Computer Science, vol 4022, Springer, Berlin (2006)

Kerremans, K., Temmerman, R., Tummers, J.:Representing multilingual and culture-specific knowledge in a VAT regulatory ontology: support from the termontography method. Proc. Workshop on Regulatory Ontologies and the Modelling of Complaint Regulations (WORM CoRe; Catana, Italy), Lecture Notes in Computer Science, Springer-Verlag, Berlin (2003)

Lauser, B., Wildemann, T., Poulos, A., Fisseha, F., Keizer, J., Katz, S.: A Comprehensive framework for building multilingual domain ontologies: creating a prototype biosecurity ontology. Proc. DC-2002 Metadata for e-Communities: Supporting Diversity and Convergence, Florence (2002)

Lauser, B., Sini, M., Liang, A., Keizer, J., Katz, S.: From AGROVOC to the agricultural ontology service / concept server an OWL model for creating ontologies in the agricultural domain. Proc International Conference on Dublin Core and Metadata Applications, Manzanillo, Colima, Mexico (2006)

Li, Y., Li, J., Zhang, D., Tang, J.: Result of ontology alignment with rimom at oaei'06. In Proc. of the International Workshop on Ontology Matching (OM- 2006), Nov. 5, 2006, Athens, Georgia, USA (2006)

Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., Peters, W.: Modelling multilinguality in ontologies. Coling 2008: Proceedings of the 22nd International Conference on Computational Linguistics, Manchester (2008)

National Library of Medicine. The Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/online_learning/OVR_001.htm (2009). Retrieved on 18 Feb 2009

Peters, W., Vossen, P., Díez-Orzas, P., Adriaens, G.: Cross-linguistic alignment of wordnets with an inter-lingual-index. Comput. Humanit. **32**, 221–251 (1998)

Rodriguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes. IEEE Trans. Knowl. Data Eng. **15**, 442–456 (2003)

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering thesauri for new applications: the AGROVOC Example. Journal of digital information, 4(4). Retrieved July 4, 2011, from http://journals.tdl.org/jodi/article/view/112 (2006)