Building a Resource of Ontological and Formal Lexical-Semantic Knowledge

Dennis Spohr
Institut für Linguistik/Romanistik, Universität Stuttgart
Stuttgart, Germany
{dennis.spohr}@ling.uni-stuttgart.de

Abstract

In this paper, we give details on our ongoing efforts to building a lexical resource that provides fine-grained lexical-semantic analyses of French verbs, in addition to a formal organisation of the ontological concepts that are used to describe them. For implementing this information, we make use of technology developed in the context of the Semantic Web, such as the Web Ontology Language OWL, Description Logic reasoners, and the Semantic Web Rule Language SWRL.

We motivate our efforts by comparing our verb analyses to those found in the French EuroWordNet (Vossen, 1998). We further show the necessity of detailed lexical-semantic knowledge – including information about presuppositions and inferences – as well as of ontological type information e.g. on permitted fillers for argument slots, for the successful completion of computational linguistic tasks. Since our resource is primarily intended for computational use, we will outline possible applications of the modelled information.

1 Introduction and Motivation

A number of large-scale lexical resources containing lexical-semantic information have been created and mapped to resources of ontological knowledge, such as WordNet and FrameNet (Fellbaum, 1998; Baker et al., 1998). Although impressive in quantitative terms, what these resources lack to a large extent is an in-depth formal lexical-semantic analysis, e.g. one that provides presuppositional and inferential information. However, this knowledge is required in order to be able to successfully perform automatic reasoning tasks such as the recognition of textual entailment.

While these resources might still serve as a solid basis for starting in-depth lexical-semantic analysis of English lexical items, there is no such resource of comparable quality for French. Although there is a French EuroWordNet (Vossen, 1998), its usability is questionable particularly because it contains a lot of inaccuracies in the description of the verbal domain. As for existing ontological resources, they tend to describe continuants (i.e. entities that are persistent through time, such as objects or organisations) with far more accuracy and detail than occurrents (i.e. entities that have temporal parts, such as events or processes).

In this paper, we will show our approach to building a resource that provides in-depth analyses of the lexical semantics of French verbs and that takes into account also presuppositional and inferential information. The lexical-semantic analyses are tightly linked to concepts in an *ontology of occurrents*. However, despite the references to large-scale lexical-semantic resources, the purpose of this paper is not to present a finished large-scale resource that is capable of directly competing with existing ones, but rather to illustrate ongoing work on principles for modelling a formal combination of syntactic, lexical-semantic, and ontological information in a single resource.

In the following section, we will introduce the necessary background and look at the extent to which existing resources could be used in the creation process. The process itself is described in detail in Section 3.

2 Background and Related Work

2.1 Formalisms

The formalisms that are used for building the resource have been developed in the field of the *Semantic Web*, a research area devoted among others to providing tools and formalisms for assigning meaning to web content (Berners-Lee et al., 2001).

In particular, we make use of the Web Ontology Language OWL (Bechhofer et al., 2004) and the Semantic Web Rule Language SWRL (Horrocks et al., 2004). While these formalisms have been described at length in the relevant literature, we will quickly summarise the main characteristics that are necessary for the comprehension of the paper.

OWL. The Web Ontology Language (Bechhofer et al., 2004) is a formalism based on the Resource Description Framework RDF[1] and can be expressed in XML syntax. Its main building blocks are classes (corresponding to one-place predicates in first-order logic), properties (two-place predicates) and individuals (instances of classes). OWL comes in three sublanguages, which differ wrt. their expressivity: OWL Lite is the least expressive sublanguage and allows for simple class definitions; OWL DL is based on description logic, a decidable fragment of first-order logic, which allows for all OWL constructs but restricts the use of some of them in order to maintain decidability of reasoning; OWL Full is the most expressive sublanguage and imposes no restrictions on the language constructs, however at the cost of decidability. For example, in OWL Full it is possible to express that a class is an instance of another class, which is disallowed in OWL DL.

SWRL. The Semantic Web Rule Language (Horrocks et al., 2004) adds expressivity to OWL in that it allows for the expression of Horn-like rules, i.e. disjunctive rules with at most one positive literal, for example

$$\neg hasFather(x,y) \vee \neg hasBrother(y,z) \vee hasUncle(x,z)$$

which is equivalent to the following rule:

$$hasFather(x,y) \wedge hasBrother(y,z) \rightarrow hasUncle(x,z)$$

SWRL can be expressed directly in OWL syntax – so the resulting documents are still OWL compliant – and the rules can be interpreted and executed by tools such as the Jess® rule engine[2].

2.2 Lexical-semantic resources and ontologies

EuroWordNet. The EuroWordNet project (Vossen, 1998) aimed at providing resources similar to Princeton WordNet (Fellbaum, 1998) for seven European languages, all of which are connected through an interlingual index (ILI) that contains a set of language-independent concepts. The ILI is linked to the so-called EuroWordNet Top Ontology, an upper-ontology-like collection of features that have been designed to describe the lexical-semantic relations in the wordnet. The French version of EuroWordNet contains roughly 8,300 verb senses and 24,500 noun senses, which are organised into 22,745 synonym sets and linked using lexical-semantic relations like hyponymy and meronymy.

In contrast to the scale of the resource in terms of covered senses, the detail of description is generally limited to taxonomic relations between synonym sets and does not include information on argument structure. However, the probably biggest drawback of the French EuroWordNet lies in its inaccuracy and even partial incorrectness, mainly wrt. to the verbal descriptions, both of which probably stem from semi-automatically translating English synsets into French (Dutoit et al., 1998). Therefore, only the noun hierarchy can be considered as a useful starting point for building other lexical resources, whereas the verb hierarchy can only provide a rough sketch as to the interpretation and organisation of the senses.

Other resources. Apart from EuroWordNet, there is no large-scale lexical resource of French that provides qualitatively adequate lexical-semantic analyses. While resources such as FrameNet and VerbNet (Baker et al., 1998; Kipper-Schuler, 2006) exist for English, none of these have been extended to French in a comparable way yet.

2.3 Ontologies

SUMO. Together with DOLCE (see below), the Suggested Upper Merged Ontology (Niles and Pease, 2001) is one of the most widely used ones in the NLP community, among others due to the fact that mappings have been created to Princeton WordNet (Niles and Pease, 2003) and the EuroWordNet ILI (Spohr, 2008a). SUMO comes with MILO, a mid-level ontology, as well as domain ontology extensions, which in total contain 20,000 terms and 70,000 axioms. While originally implemented in SUO-KIF – a formalism intended as first-order language – SUMO has also been translated to OWL Full, with the attempt to preserve as much as possible of the original axiomatisation.

[1] http://www.w3.org/RDF/

[2] http://www.jessrules.com/

Despite its quantitative size and degree of formalisation, SUMO has been criticised primarily wrt. the usability of its axiomatisations, since they are questionable from a modelling perspective (e.g. instances being concepts at the same time and relations being modelled as concepts). Moreover, SUMO seems to lack a clear theoretical basis, as it adopts ideas from different ontological theories (Sonntag et al., 2007).

DOLCE. The Descriptive Ontology for Linguistic and Cognitive Engineering (Gangemi et al., 2003a) is an upper-level ontology that has been designed with a strongly cognitive bias. Its classes and the relations among them have been implemented with the OntoClean methodology (Guarino and Welty, 2002), which gives the resource a formally and theoretically more solid basis than e.g. SUMO. As was mentioned above, DOLCE has also been mapped to Princeton WordNet (Gangemi et al., 2003b).

DOLCE is the first reference module of the WonderWeb library of foundational ontologies, and it has a number of extensions (e.g. an ontology of information objects). In total, DOLCE and its extensions comprise roughly 200 classes and 300 properties, and they are available as OWL versions.

Next to this version of DOLCE, which is called *DOLCE-Lite-Plus*, there exists a version called *DOLCE-Ultralite* (DUL), which uses friendly names for classes and properties and simple class restrictions.[3] For these reasons, and since DUL is – as DOLCE-Lite-Plus – expressed in OWL DL, it provides a solid formal basis for the definition of a lexical-semantic and ontological resource. In total, DUL contains roughly 200 classes and 130 properties.

3 Creation and Computational Use of the Resource

In the following, we will discuss the different steps in the process of building the resource. The manual analysis that precedes the other ones will be omitted here since it has been discussed at length in (Martin et al., to appear). However, it is important to notice that at the end of this analysis step, we have obtained a formal lexical-semantic representation of different senses of a verb that contains information about presuppositions and inferences, in addition to information about sense-specific restrictions on the ontological type of argument slot fillers (e.g. "the subject has to be human" or "a directional prepositional phrase has to be present").

3.1 Interfaces between syntactic, ontological and lexical-semantic knowledge

In this section, we will explain how we model the knowledge obtained from the manual analysis, on the one hand in the form of a kind of "lexical entry" for the different senses, on the other hand in the form of ontological concepts and inference rules.

Ontological argument restrictions in the lexicon. On the basis of the above analysis, we create a small subhierarchy of classes in our lexicon, corresponding to the senses of a verb. The classes are organised hierarchically (as shown in Figure 1) in order to be able to express generalisations that hold for more than one sense, and in order to be able to complete reasoning tasks such as "is the occurrence of *pousser* in this sentence a physical sense of *pousser*?".

pousser
- *pousser_figurative*
  - *pousser_conceptual*
    - $pousser\_conceptual_1$
    - $pousser\_conceptual_2$
    - ...
  - *pousser_psychological*
    - $pousser\_psychological_1$
    - ...
- *pousser_physical*
- ...

Figure 1: Hierarchy of senses of *pousser*

Each of the "leaf classes" (e.g. $pousser\_conceptual_1$) represents a specific configuration of syntactic and ontological parameters, which are modelled as necessary and sufficient conditions on the definition of the respective class. These axioms are the result of expressing the findings and intuitions wrt. the ontological type of the arguments in the manual analysis step in terms of DOLCE-Ultralite concepts. Figure 2 below shows such a configuration for one of the conceptual senses of *pousser*.

The formalisation is to be interpreted as follows: in order to be classified as an instance of

[3] http://wiki.loa-cnr.it/index.php/LoaWiki:DOLCE-UltraLite

$pousser\_conceptual_1 \equiv pousser$
$\exists\ subj\ (\exists\ canDenote\ dul{:}Organism)$
$\forall\ subj\ (\exists\ canDenote\ dul{:}Organism)$
$\exists\ obj\ (\exists\ canDenote\ dul{:}Abstract)$
$\forall\ obj\ (\exists\ canDenote\ dul{:}Abstract)$
$\geq\ 3\ arg\ \ owl{:}Thing$

Figure 2: Axiomatisation of $pousser\_conceptual_1$

$pousser\_conceptual_1$, it is both necessary and sufficient to be an instance of *pousser*, with a subject that can denote an organism, with a direct object that can denote something abstract, and with at least one more argument (i.e. the number of arguments is at least 3; *owl:Thing* just refers to "any kind of entity"). The predicate *canDenote* used in the formalisation captures the polysemy of the nominal argument, since the classes that represent nouns contain as axioms the ontological concepts they can denote, such as e.g. the class $faim_1$ with the axiom $\exists$ *canDenote dul:SocialObjectAttribute*. So in other words, the object part of the example above states that the value of the *obj* property of *pousser* has to be an instance of a class that can denote something abstract (i.e. *dul:Abstract* or any of its subclasses). An example of an instance of this sense of *pousser* is given in sentence 1 below.

(1) *Pierre a poussé ma faim*
Pierre has pushed my hunger
jusqu'à la rage.
to the point of fury.

As can be seen in the figure, we have implemented a very tight link between ontological and syntactic information. In addition to this, we have a further link from the syntax to the ontological and formal lexical-semantic analysis, which will be illustrated in the following.

Inference rules. In order to model the inferences triggered by the syntactic configurations shown above, a formalism that goes beyond the expressivity of OWL is needed, e.g. to be able to make assertions about the entities involved. For this we make use of SWRL rules that contain a specific syntactic configuration in the rule body (e.g. $pousser(?e) \wedge subj(?e,?x) \wedge obj(?e,?y)$) and a resulting lexical-semantic output configuration in the rule head (e.g. $PUSHING(?e) \wedge agent(?e,?x) \wedge VECTOR(?v) \wedge source(?e,?v)\ldots$). Such a rule is interpreted for example as "if we have an instance e of *pousser* with subject x and object y, then e is also an instance of a *PUSHING*-event, with x as agent and a vector v as source ...". Thus, rules implement a crucial link between the syntax on the one hand, and lexical-semantic and ontological knowledge on the other.

Ontology of occurrents. As can be seen in the rule excerpt above, we make use of other ontological concepts in addition to the ones defined in DUL, such as *PUSHING* and *VECTOR*. Taxonomically, they are located below the DUL concepts in the hierarchy, as they represent more specific cases of the ones defined there, e.g. *PUSHING* as a more specific kind of *Action*. The aim of this *ontology of occurrents* is to also assign axiomatic definitions and inference rules to the concepts therein, in order to generalise conceptual properties over specific lexical realisations, i.e. verb senses. This ontology is still work in progress, and since we intend to design it according to the OntoClean principles, we have used DUL to sort of "prestructure" our concept hierarchy. However, defining essential and rigid properties or identity criteria of occurrents is an entire topic of its own, and will be part of future research.

3.2 Computational use

In the following paragraphs, we will briefly explain how the resource can be used for automatic word-sense disambiguation and calculation of inferences.

Disambiguation of verbs in context. The primary factors that can be used for the disambiguation of verb senses is the ontological type of the syntactic arguments. As was shown in Figure 2 above, these are modelled as necessary and sufficient conditions in the respective class definition.

For disambiguating a sentence like the one in (1), we would first assume syntactic input that provides at least information about the predicate (*pousser*), its arguments (e.g. *subject* = *Pierre*, *object* = *faim* etc.) as well as the tense used (in this case *passé composé*). The fillers of the argument slots are then looked up in selectional preference lists of the respective predicate (Spohr, 2008b), which contain information about the most probable ontological types per argument slot, and the sense of the noun whose ontological type scores highest is selected and asserted in the resource.

For example, after having selected a sense of *faim*, we assert an individual *x* as an instance of the class $faim_1$ and link it to the predicate by means of the *subj* relation, i.e. *subj(x)*. On the basis of (i) the syntactic configuration, (ii) the necessary and sufficient conditions in the classes for *pousser*, and (iii) the sense selection for the nominal arguments, a description logic reasoner (e.g. Pellet; (Sirin et al., 2007)) is run and infers a sense of the predicate *pousser* that has been used in this particular sentence.

Calculation of inferences. Once a sense has been selected by the reasoner, the system can execute the SWRL rules that have been defined for the respective senses in order to calculate the inferences that are licensed on the basis of the previous sense selection. As was mentioned in Section 3.1, the appropriateness of a rule is further determined by the syntactic context in which the verbal predicate has been used, and which has to match with the one stated in the rule body. The new statements that result from the rule execution are then asserted in the resource. They represent the logical form of the input sentence, based on the ontologically enriched manual lexical-semantic analysis. This information, which is directly encoded in OWL, can then further be made available to other applications.

3.3 Current state of and future plans

As was already mentioned in the introduction, the resource is not in a state of being applied to real-life tasks. The lexical-semantic analysis of verbs as well as the definition of the ontology are still work in progress, and the current size in terms of senses covered is very small. Nonetheless, sample tests on selected corpus sentences have been able to serve as a proof of concept for the rich formalisation of verbs as being done in our project. Therefore, with the formal principles of modelling lexical-semantic and ontological information defined, we intend to tackle the quantitative size of the resource in the future.

4 Conclusion

In this paper, we have provided details on the process of building a lexical resource of French that contains a high level of detail wrt. the lexical-semantic and ontological analysis of the verbal domain, with focus on the interplay between syntactic, lexical-semantic and ontological information. In addition to motivating the necessity of a high level of detail in the modelling of this knowledge, we have presented ongoing efforts in designing an ontology of occurrents and, finally, outlined the potential of the resulting resource for use in computational scenarios.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the joint COLING/ACL 1998*, Montreal, Canada.

Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2004. OWL Web Ontology Language Reference. W3C Recommendation. `http://www.w3.org/TR/owl-ref/`.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–43.

Dominique Dutoit, Laurent Catherin, and Andreas Wagner. 1998. Specification of German & French WNs. EuroWordNet (LE-8328) Deliverable: 2D002.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003a. Sweetening WordNet with DOLCE. *AI Magazine*, 24(3):13–24.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003b. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In *Proceedings of ODBASE*, Catania, Italy. Springer.

Nicola Guarino and Christopher Welty. 2002. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2):61–65.

Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission. `http://www.w3.org/Submission/SWRL/`.

Karin Kipper-Schuler. 2006. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia.

Fabienne Martin, Dennis Spohr, and Achim Stein. (to appear). Representing a Resource of Formal

Lexical-Semantic Descriptions in the Web Ontology Language. *GSCL Forum – Special Issue on Lexical-Semantic and Ontological Resources*.

Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS 2001)*, Ogunquit, ME.

Ian Niles and Adam Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03)*, Las Vegas, NV.

Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. 2007. Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 5(2).

Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf, Norbert Pfleger, Massimo Romanelli, and Norbert Reithinger. 2007. SmartWeb Handheld – Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In Thomas S. Huang, Anton Nijholt, Maja Pantic, and Alex Pentland, editors, *Artificial Intelligence for Human Computing*, volume 4451 of *Lecture Notes in Artificial Intelligence*, pages 272–295. Springer, Heidelberg.

Dennis Spohr. 2008a. A General Methodology for Mapping EuroWordNets to the Suggested Upper Merged Ontology. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Dennis Spohr. 2008b. Extraction of Selectional Preferences for French using a Mapping from EuroWordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 4th Global WordNet Conference*, Szeged, Hungary.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.