# Ontology Learning from Text Using Automatic Ontological-Semantic Text Annotation and the Web as the Corpus

**Jesse English and Sergei Nirenburg**

Institute for Language and Information Technologies
University of Maryland, Baltimore County
Baltimore, MD 21250, USA

## Abstract

We present initial experimental results of an approach to learning ontological concepts from text. For each word to be learned, our system a) creates a corpus of sentences, derived from the web, containing this word; b) automatically semantically annotates the corpus using the OntoSem semantic analyzer; c) creates a candidate new concept by collating semantic information from annotated sentences; and d) finds in the existing ontology concept(s) "closest" to the candidate. In the long term, our approach is intended to support the continual mutual bootstrapping of the learner and the semantic analyzer as a solution to the knowledge acquisition bottleneck problem in AI.

## 1. Introduction

Automating knowledge acquisition for use in automatic reasoning systems in a variety of applications has long been recognized as the Holy Grail of AI. In recent years, work in this area has gained momentum as an application of machine learning for rapid knowledge formation, as a requirement for the success of the Semantic Web enterprise, as a means of facilitating the development of ontologies and as a step toward attaining the ultimate goal of teaching computers to learn from reading text.

The long-term goal of our ongoing research is indeed learning by reading. Specifically, we are working toward creating a system (an intelligent agent) that will be able to extract from text formal representations ready for use in automatic reasoning systems. These structures will reflect both instances and types of events, objects, relations and agents' attitudes in the real world. The reasoning that such agents will be able to perform will support both general problem solving and, specifically, knowledge-based NLP, that is, the very process through which the agent learns from text.

We model learning by reading as the process whereby an agent:

- analyzes (reads) a text and generates text meaning representations ready for use in a reasoning system (either in a task-oriented situation in an application or in a dedicated learning mode) and, when a certain word is found that it (the agent) does not know (that is, it is not in the agent's lexicon),
- undertakes to learn the meaning and the syntactic, morphological and collocational features of the word automatically,
- adds the newly learned word to the lexicon and
- continues the original reading process.

This process presupposes the existence of a text analyzer capable of producing structures of requisite depth. The learning process, thus, benefits from the existence of the analyzer and at the same time benefits the analyzer by enriching its static knowledge resources through learning by reading. This mutual bootstrapping methodology is very promising and will be used in our project both in an unsupervised and a supervised setting. The former will accept candidate knowledge elements learned by the system and seek to evaluate the quality of the additions to the knowledge resources after a particular number of new knowledge elements are learned. The latter will introduce a human validation/correction step and thus ensure that the quality of newly acquired knowledge is commensurate with that of the system's knowledge resources at the beginning of the learning process. Thus, our research will be both evaluating the efficiency of our learning methods and practically contribute to the growth of the knowledge resources of our analyzer. Indeed, from the point of view of practical utility, it is safe to assume that the level of automation of high-quality knowledge acquisition will increase gradually, with humans playing a role in this process for some time to come.

In this paper we describe our initial results on a learning-by-reading experiment of the type described above, concentrating on learning the meaning of unknown lexical units. The paper is organized as follows: first, we briefly describe the technology underlying the experiment, namely, the OntoSem text analyzer and its knowledge resources. Next, we present the experimental set-up, describe the processing that was involved in the experiment, present initial results, evaluate and discuss them. Finally, we describe planned extensions and modifications to our methodology.

## 2. OntoSem

OntoSem (the implementation of the theory of Ontological Semantics; Nirenburg and Raskin 2004) is a text-processing environment that takes as input unrestricted raw text and carries out preprocessing, morphological analysis, syntactic analysis, and semantic analysis, with the results of semantic analysis represented as formal text-meaning representations (TMRs) that can then be used as the basis for many applications. TMRs have been used as the substrate for question-answering (e.g., Beale et al. 2004), machine translation (e.g., Beale et al. 1995) and knowledge extraction, and were also used as the basis for reasoning in the question-answering system AQUA, where they supplied knowledge to showcase temporal reasoning capabilities of JTP (Fikes et al., 2003). Text analysis relies on extensive static knowledge resources:

- The OntoSem language-independent **ontology**, which currently contains around 8,500 concepts, each of which is described by an average of 16 properties. The ontology is populated by concepts that we expect to be relevant cross-linguistically. The current experiment was run on a subset of the ontology containing about 6,000 concepts.
- An OntoSem **lexicon** whose entries contain syntactic and semantic information (linked through variables) as well as calls for procedural semantic routines when necessary. The semantic zone of an entry most frequently refers to ontological concepts, either directly or with property-based modifications, but can also describe word meaning extra-ontologically, for example, in terms of modality, aspect or time (see McShane and Nirenburg 2005 for in-depth discussion of the lexicon/ontology connection). The current English lexicon contains approximately 30,000 senses, including most closed-class items and many of the most frequent and polysemous verbs, as selected through corpus analysis. The base lexicon is expanded at runtime using an inventory of lexical (e.g., derivational-morphological) rules.
- An **onomasticon**, or lexicon of proper names, which contains approximately 350,000 entries.
- A **fact repository**, which contains "remembered instances" of ontological concepts (e.g., SPEECH-ACT-3366 is the 3366$^{th}$ instantiation of the concept SPEECH-ACT in the memory of a text-processing agent). The fact repository is not used in the current experiment but will provide valuable semantically-annotated context information for future experiments.
- The OntoSem syntactic-semantic **analyzer**, which performs preprocessing (tokenization, named-entity and acronym recognition, etc.), morphological, syntactic and semantic analysis, and the creation of TMRs.
- The TMR language, which is the **metalanguage** for representing text meaning (we have recently developed a converter between this custom language and OWL, see Java et al. 2005).

OntoSem knowledge resources have been acquired by trained acquirers using a broad variety of efficiency-enhancing tools – graphical editors, enhanced search facilities, capabilities of automatically acquiring knowledge for classes of entities on the basis of manually acquired knowledge for a single representative of the class, etc. OntoSem's DEKADE environment (see McShane et al. 2005) facilitates both knowledge acquisition and semi-automatic creation of "gold standard" TMRs, which can be also viewed as deep semantic text annotation. A high-level view of OntoSem text processing is shown in Figure 1.

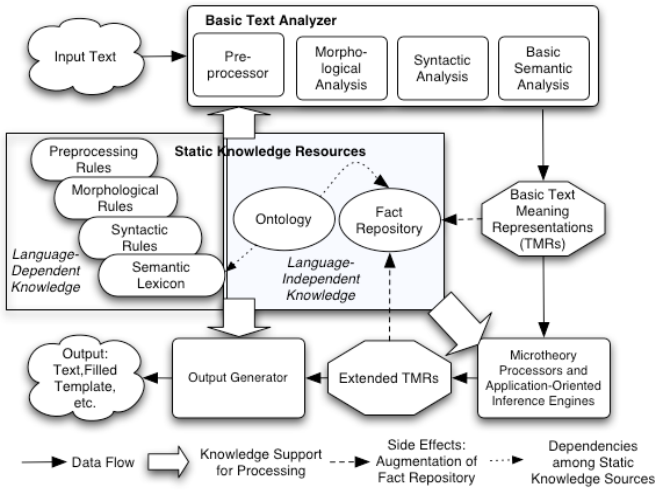


Figure 1. A High-Level View of OntoSem.

## 3. The Learning Experiment

We concentrate on learning the meaning of unknown words. One simplifying assumption we make at this time is that the meaning of the candidate will be expressed as a univocal mapping into an ontological concept (in general, SEM-STRUC zones of OntoSem lexicon entries can add local constraints to ontological concepts in terms of which the meaning of the lexical unit is described, thus making such lexicon entries unnamed ontological concepts). As a result, at this time, the experiment is effectively constrained to learning ontological concepts.

Ontology learning as a field concerns itself at this time with learning terms, (multilingual) synonyms, concepts, taxonomies (by far the most popular topic), relations and rules and axioms (Buitelaar et al. 2005). The methods involved include different combinations of linguistic (knowledge-based) and statistical methods but mostly the latter. Among the linguistic tools used for this

purpose Buitelaar et al., *ibid*., list, in order of increasing sophistication, tokenization, part of speech and semantic type tagging, morphological analysis, phrase recognition, (syntactic) dependency structure determination and discourse analysis. It is notable that this list does not include a tool that would be centrally important for learning by reading – an analyzer that creates disambiguated semantic dependency structures for sentences and texts, in which relations between elements are ontological and not syntactic and go well beyond the taxonomic subsumption relations. Work on extracting small subsets of such relations using largely statistical means has been reported (e.g., Charniak and Berland 1999 for meronymy, Cimiano and Wenderoth 2005 for the qualia of the generative lexicon approach (Pustejovsky 1995), causation (Girju 2002), among others). OntoSem, however, addresses the task of extracting knowledge about a large set of such relations using encoded knowledge as heuristics (cf. work by, e.g., Clark and Weir 2002 that uses essentially statistical methods for estimating selectional restrictions).

Among the sources of knowledge acquisition are machine-readable dictionaries (e.g., Nichols et al. 2006), thesauri (e.g., Navigli and Velardi 2006), as well as text (e.g., Ogata and Collier 2004, Buitelaar et al. 2004, Cimiano et al. 2005). Our experiment uses open text but can be extended to treating MRDs and thesauri as special types of texts.

Most extant ontology learning methods operate at the level of textual strings, using sophisticated statistical analysis and clustering algorithms (optionally augmented by relatively shallow linguistic analysis) and thus requiring annotated training corpora. Our approach, by contrast, relies on a dynamically generated corpus of knowledge structures (TMRs) written in an ontological metalanguage, obtained through the operation of OntoSem, which relies on deep linguistic analysis strengthened by statistical algorithms operating over the ontology and the nascent TMRs. At present, the quality of automatically generated TMRs is not optimal. The plan is to improve the quality of TMRs through learning new ontological and lexical knowledge using the current state of OntoSem, with or without using human validators/editors to "goldenize" system-produced TMRs. Thus, an important empirical question that our experimentation seeks to answer is how much (or how little) human intervention is needed at any time in the knowledge acquisition process to sustain the continuous growth and improvement of the knowledge resources while maintaining and enhancing their quality.

Our initial experiment starts with selecting words whose meaning will be learned by the system. Next, we develop a corpus of sentences containing this word and use OntoSem to generate their TMRs. OntoSem degrades gracefully in the face of unexpected input. It is capable of semantically analyzing sentences with a small number of unknown words by assuming that the unknown word's meaning corresponds directly to a non-existent ontological concept and then applying relevant constraints listed in the knowledge about words syntactically connected with the unknown word to hypothesize the constraints on the latter. Note that, unlike the case when all the words are known and, therefore, constraints are mutually **matched** (this is a major mechanism for lexical and semantic dependency disambiguation, the core process of semantic analysis), in the case of an unknown word the constraints (for example, selectional restrictions) must be applied unidirectionally.

From the TMRs containing the new candidate concept, our system collects both the inventory of relations and attributes attested for the candidate concept and the inventory of values of these relations and attributes. After the candidate concept is thus "assembled," we compare it with the concepts in the existing ontology to find the most appropriate position(s) for it in the multiple-inheritance hierarchy that organizes the OntoSem ontology. To facilitate the evaluation of our experimental results, we have also taken some of the existing entries (and the corresponding ontological concepts) out of the OntoSem knowledge resources, run the experiment and then compared the automatically generated new concepts with the original ones. In what follows, we describe this process in somewhat greater detail.

## 4. The Experiment

We are using the web as our corpus. Specifically, the learner uses Google's SOAP Search API (http://www.google.com/apis/index.html), which returns a list of websites containing the word. The content of a specified number of these sites is retrieved. This number can be adjusted if needed to extract a larger corpus. Next, an HTMLParser (http://htmlparser.sourceforge.net/) is used to strip out tags and unwanted markup, yielding raw text. At the next step, text is divided into sentences (using a module of the OntoSem analyzer), and those sentences that do not contain the search word are discarded.

The remaining sentences are processed using the OntoSem analyzer that carries out morphological, syntactic and semantic analysis. The output from OntoSem is a list of TMRs containing instances of a candidate concept corresponding to the unknown word and instances of a variety of relations in which this concept participates, such as a case role that relates it to an event-type concept instance. Additionally, the TMR may contain some values for attributes (unary properties) of the candidate concept. At this point in the process there is an option to include human assistance to produce gold-standard TMRs from the system's results. Though we have a tool, DEKADE, available for this purpose, we did not at this time use human involvement in the process.

The list of properties returned from any one TMR is likely to be small (because only a few will be referred to in a single sentence). The system then has to collate the

knowledge extracted from processing individual sentences. Given a list of TMRs, the learner searches through each one, finding all properties associated with the instances of the candidate concept, and collating them into a single frame for the corresponding candidate ontological concept. When collating the values of each property of the candidate concept, the system filters out weaker constraints if stronger constraints have been attested. For example, if among the fillers of the AGENT-OF property of the candidate concept the system finds READ, ACTIVE-COGNITIVE-EVENT and MENTAL-EVENT, it will retain only READ because it is a descendent of the other two concepts in the ontology. The weaker constraints can, in principle be retained in the OntoSem ontology because the latter uses multi-level constraints to support robust disambiguation processes. Technically, the constraint read may appear in the SEM facet of the property AGENT-OF in the candidate concept while MENTAL-EVENT may appear as the filler of the RELAXABLE-TO facet of the same property.

The final step in our current experiment is to find existing ontological concepts that are most similar to the candidate concept, with the idea of suggesting a place for the new concept in the multiple-inheritance hierarchical organization of the OntoSem ontology. In general, there are three distinct outcomes:

1. The candidate concept is subsumed by an existing concept (meaning that the original unknown word is a synonym of an existing lexical entry).
2. The candidate concept is similar to an existing concept C, in which case the system will create a lexicon entry for the original unknown word and in the SEM-STRUC zone of this entry insert a reference to C, with further local constraints added to reflect the differences between the candidate concept and concept C.
3. The candidate concept is sufficiently dissimilar from existing concepts, and should be added to the ontology.

The method of using the results of learning sketched above will be included (with an optional human validation step) in OntoSem's knowledge acquisition environment. We will report about our work on automatic creation of OntoSem lexicon entries elsewhere. Here, we describe how we create and evaluate ontology concepts. We used two strategies: using genuinely new words (those not in the OntoSem lexicon) and using existing words and removing them from the lexicon (and concepts used to describe their meaning from the ontology) before the corresponding system run. In the latter strategy, we thus had a gold standard concept against which to compare the candidate. In the former strategy, we picked a concept in the existing ontology that we thought would be the most appropriate one to use in the description of the meaning of the new word (whether directly, through constrained lexical mapping or creation of a new concept).

## 5. Results and Evaluation

We present the results of the learning of four words – *hobbit* and *pundit* were entirely new to the system and *CEO* (and its corresponding concept PRESIDENT-CORPORATION) and *song* (and its corresponding concept SONG) were temporarily deleted from the OntoSem knowledge resources.

For *hobbit*, we selected HUMAN as the closest ontological concept. Table 1 shows a comparison of the selected properties of automatically generated candidate concept for *hobbit*, HOBBIT, and the existing concept HUMAN.

| Ontological Property | Values in HUMAN | Values in HOBBIT |
|---|---|---|
| AGENT-OF | LIVE<br>CREATE-ARTIFACT<br>ELECT<br>READ | LIVE<br>CREATE-ARTIFACT<br>ELECT |
| THEME-OF | RESCUE<br>MARRY<br>KILL | RESCUE<br>KILL |
| HAS-OBJ-AS-PART | HEAD | n/a |

Table 1: Comparison of selected properties of HUMAN to automatically generated candidate concept HOBBIT for the word *hobbit*.

Present in many of the automatically generated concepts were a relatively high proportion of properties labeled RELATION, which means that the system was not able to determine a more precise link (that is, a narrower-defined property) connecting the candidate concept with the filler of a RELATION instance. The OntoSem ontology uses approximately 200 specialized relations to characterize objects (the subtree of properties also contains a comparable number of attributes, and one-place predicates). The OntoSem analyzer uses the concept RELATION when it determines that two concept instances are related but lacks heuristics to specify what the specific relation it is. This over-generalized output is the price we pay for making sure that the analyzer does not break on unexpected input or due to insufficient quality of the existing knowledge resources or decision heuristics. Even though the connection on RELATION is relatively underspecified, at present we keep this information and use it alongside other constraints in determining the distance between the candidate concept and other concepts in the ontology.

Our evaluation is based on measuring the ontological distance between the candidate concept and all the other concepts in the ontology and then determining a) the difference in the distances between the candidate concept and the automatically derived closest concept and to the designated hand-picked closest concept; and b) the

rank of the hand-picked concept in the sorted list of closest concepts. Table 2 presents the results of our initial experiment.

Distances between ontological concepts are measured using the OntoSearch algorithm (Onyshkevych 1997). OntoSearch finds the "best" ontological path (chain of relations) between any two concepts and calculates the weight (score) of each path, which reflects the strength of the association between two concepts. The cumulative score for a path is a function of its length and of the cost of traversing a particular relation link. For example, subsumption links (IS-A and SUBCLASSES) are less costly to traverse than, say, causal links. The individual link traversal costs in OntoSearch were trained using simulated annealing on a representative subset of OntoSem ontological relations. OntoSearch has been used to provide statistics-based heuristics to supplement static knowledge resources during the operation of the OntoSem text analyzer. For example, to help disambiguate the input *The doctor performed the operation*, Ontosearch examines the ontological connection between the non-title sense of *doctor* and the two senses of *operation*: PERFORM-SURGERY and MILITARY-ACTIVITY and returns the following link:

ONTOSEARCH(DOCTOR, PERFORM-SURGERY) → 0.8
DOCTOR AGENT-OF PERFORM-SURGERY

(In other words, even though the ontology may not overtly contain the information that a doctor performs surgeries, this information is virtually there and can be derived by the analyzer using OntoSearch.) The OntoSearch score for the DOCTOR/MILITARY-ACTIVITY relationship is much lower, so that the PERFORM-SURGERY sense is preferred.

As an evaluative tool, OntoSearch allows us to show both a correctness metric, as well as a measure of improvement in results given different sizes of a dynamically produced relevant text corpus. As the ultimate goal of this research is developing a learning-by-reading capability, this provides data for testing the hypothesis: whether indeed the more the learner reads, the more it useful knowledge it obtains. The corresponding data for our experiment is summarized in Figure 2. Determining saturation points in learning will help the efficiency of learning by reading on complete texts.

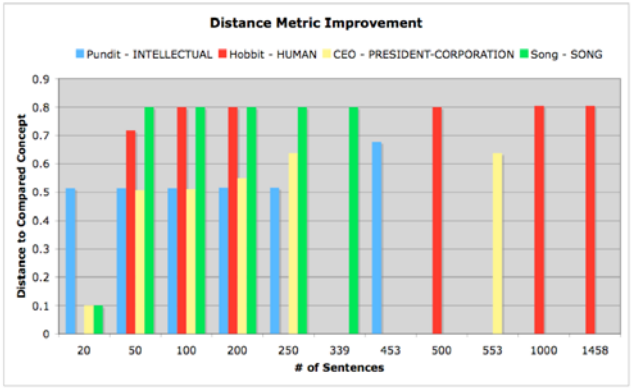


Figure 2: Improvement of generated concept vs. target concept over growing corpuses.

## 5. Discussion and Future Work

We view the results obtained so far in our work as baseline-level results because the current experiment was conducted using existing technology and knowledge resources that were not (yet) tuned for the task of learning by reading. There are multiple avenues of improvement. Thus, the quality of the OntoSem knowledge resources can and will be improved. We are running a parallel experiment on using a corpus empirically derived from the web (using the same methodology as in this paper) and the existing OntoSem knowledge resources to empirically validate and improve the fillers of properties of existing ontological concepts. As a result of this work, we expect the analyzer to start generating fewer properties marked simply RELATION and include more specific, information-rich properties.

At present, since learning precision is more important than recall (especially when dealing with an open-ended corpus, such as the web), sentences which OntoSem fails to analyze completely are simply discarded. We have several ways of reducing the number of discards, notably, moving down to the clause level of analysis from the sentence level.

In the current experiment we used a similarity measure, OntoSearch, that was originally developed to support ambiguity resolution. We will develop a

| Word | Best Match | Selected Match | Difference | Rank | Percentile |
|---|---|---|---|---|---|
| pundit | TELEVISION, CITIZEN, HUMAN, (and 12 more). 0.800 | INTELLECTUAL 0.679 | 0.121 | 210/~6000 | 3.5% |
| ceo | EVENT 0.900 | PRESIDENT-CORPORATION 0.638 | 0.262 | >500/~6000 | >8.3% |
| hobbit | PUBLISH 0.900 | HUMAN 0.806 | 0.094 | 18/~6000 | 0.3% |
| song | WORD, RECORD-TEXT, OBJECT (and 8 more). 0.800 | SONG 0.800 | 0.000 | 12/~6000 | 0.2% |

Table 2: Initial results for processed words, their best matches, and the rank of the target.

similarity measure dedicated to learning by reading. We intend to test several of the current methods, for instance, those discussed in Curran and Moens (2002).

In this experiment we used only the concepts directly listed in the ontology. The semantic representation substrate of OntoSem actually also includes the structures in the SEM-STRUC zones of those lexicon entries that contain constrained mappings to ontological concepts. These structures effectively have the status of unnamed ontological concepts indexed through word senses (the reason for not including them in the ontology is due to the desire to separate truly language-independent meanings which, thus, belong in an ontology from language-specific "packaging" of meaning. When constructed in this way, the ontology can serve as the interlingua for machine translation, which many years ago was the initial intended application of OntoSem). In follow-up experiments, we will use the content of the SEM-STRUC zones of the English lexicon alongside the ontology concepts to compare against the automatically generated concepts.

The OntoSem analyzer was used "as is" and therefore used a powerful constraint relaxation mechanism to maintain robustness in the face of unexpected input. We will experiment with different relaxation settings to tune this relaxation capability to the needs of learning, where being able to use tighter (more information-laden) constraints on the unknown word in question is more valuable than returning a complete (though underspecified) analysis for a sentence.

The preprocessor used during the text analysis stage of the current experiment did not mark the cases where *hobbit* was used as a part of the title of the book, and this is reflected in the results presented in table 1. We will introduce better named-entity filtering in the next version of the system.

## References


Beale, S., S. Nirenburg, K. Mahesh. Semantic Analysis in the Mikrokosmos Machine Translation Project. In: Proceedings of the 2nd Symposium on Natural Language Processing, pp. 297-307, 1995.

Beale, S., B. Lavoie, M. McShane, S. Nirenburg, T. Korelsky. (2004). Question Answering Using Ontological Semantics. *Proceedings of ACL-2004 Workshop on Text Meaning and Interpretation*. Barcelona, Spain.

Buitelaar, P., S. Handschuh and B. Magnini (eds.) 2004. Proceedings of the ECAI-2004 Workshop on Ontology Learning and Population (OLP). Valencia, Spain, August.

Buitelaar, P. P. Cimiano, M. Grobelnik, M. Sintek. 2005. Ontology Learning from Text. Tutorial at ECML/PKDD, Porto, Portugal, October.

Buitelaar, P., P. Cimiano, B. Loos. (eds.). OLP-06. Proceedings of the 2nd Workshop on Ontology Learning and Population. COLING/ACL 2006, Sydney, Australia.

Charniak, E., M. Berland. Finding parts in very large corpora. In: Proceedings of the 37th Annual Meeting of the ACL, pp. 57-64, 1999.

Cimiano, P., G. Ladwig and S. Staab. 2005. Gimme' the context: Context-driven automatic semantic annotation with c-pankow. Proc. 14th WWW. ACM, 2005.

Cimiano, P., J. Wenderoth, Automatically Learning Qualia Structures from the Web. In: Proceedings of the ACL Workshop on Deep Lexical Acquisition, pp. 28-37, 2005.

Clark, S., D.J. Weir. Class-Based Probability Estimation Using a Semantic Hierarchy. Computational Linguistics, 28(2), pp. 187-206, 2002.

Curran, J., M. Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pages 59-66, Philadelphia, PA, USA.

Fikes, R., J. Jenkins, G. Frank. *JTP: A System Architecture and Component Library for Hybrid Reasoning*. Technical Report KSL-03-01, Knowledge Systems Laboratory, Stanford University, Stanford, CA, USA, 2003.

Girju, R., D. Moldovan, Text Mining for Causal Relations, In: Proceedings of the FLAIRS Conference, pp. 360-364, 2002.

McShane, M., S. Nirenburg, S. Beale. 2005. An NLP Lexicon as a Largely Language Independent Resource. Machine Translation 19(2): 139-173.

McShane, M., S. Nirenburg, S. Beale, T. O'Hara. 2005. Semantically Rich Human-aided Machine Annotation. Proceedings the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, ACL-05, Ann Arbor, June 2005, pp. 68-75.

Navigli, R., P. Velardi. 2006. *Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain*. Proceedings of OLP-06.

Nichols, E., F. Bond, T. Tanaka, F. Sanae and D. Flickinger. 2006. *Multilingual Ontology Acquisition from Multiple MRDs.* Proceedings of OLP-06.

Nirenburg, S., V. Raskin. Ontological Semantics. SERIES: Language, Speech, and Communication, MIT Press, 2004.

Ogata, N., N. Collier. 2004. Ontology Express: Non-Monotonic Learning of Domain Ontologies from Text. Proceedings of OLP.

Onyshkevych, B. 1997. Ontosearch: Using an ontology as a search space for knowledge based text processing. Unpublished PhD Dissertation. Carnegie Mellon University.

Pustejovsky, J. 1995. *The Generative Lexicon.* Cambridge/London: MIT Press.