Constructing a Corpus-based Ontology using Model Bias [*]

A. Rumshisky, P. Hanks, C. Havasi, J. Pustejovsky
Department of Computer Science Brandeis University
Waltham, MA 02453
{arum,patrick,havasi,jamesp}@cs.brandeis.edu

Abstract

Recent work in lexical resource construction has recognized the importance of contextualizing the knowledge in existing resources and ontologies with information derived from text corpora. This paper describes the integration of a corpus-based lexical acquisition process with a large, linguistically motivated lexical ontology. This semi-automatic bootstrapping process is used to produce refinements, additions, and modifications to the type specifications for the arguments to predicates in this ontology. In addition, the procedure is used to verify and modify the lexical extensions of the entity types in the ontology.

Introduction

The construction of lexical resources for natural language processing tasks is one that is arguably best done with explicit reference to the linguistic phenomena and constraints in a particular language. This is the strategy adopted by the generative lexicon based model used in the SIMPLE project, for example (Busa, Calzolari, & Lenci 2001). Databases and ontologies of various size and scope have been built on these principles, including the Brandeis Semantic Ontology (BSO) (Pustejovsky *et al.* 2005). These lexical databases have proven useful to the field and have their own merit. But even with the care and linguistic sensitivity taken in the design of such resources, they suffer from not being supported by evidence from a language corpus and contextualized to specific linguistic usage. Although it is important to impose a theoretical "model bias" on the structure of a knowledge base, recent work in lexical resource construction has recognized the importance of contextualizing this knowledge with information derived from text corpora.

To remedy that issue, we have been working on a project to capture the contextual environment for predicators in English. This is being done by the construction of a *context dictionary*, using a methodology called *corpus pattern analysis* (CPA) (Hanks & Pustejovsky 2005). In this paper we describe our efforts at merging these two research efforts in order to exploit contextual information derived for predicators in CPA in order to enrich selectional preference and typing environment in the BSO's ontology.

Contextualizing Lexical Resources

Many existing lexical resources in the community suffer from insufficient grounding in real corpus data. This is not surprising, since knowledge engineering for lexical resources is currently still a labor-intensive process; as a result, lexical databases such as WordNet are widely accepted as viable resources in the NLP community, in part due to their wide coverage. Nevertheless, the lexical entries in both MRDs and lexical databases often suffer from lack of explicit and well-founded contextualization.

WordNet

The great merit of WordNet is that it is a full inventory of English words. WordNet assigns words to "synsets" (synonym sets), which are equated with "senses". But in many cases WordNet's senses are indistinguishable from one another by any syntactic, syntagmatic, or semantic criteria. For example, in WordNet 2.1 the verb *write* is said to have 10 senses, the first four of which are listed below:

1. write, compose, pen, indite – (produce a literary work; "She composed a poem"; "He wrote four novels")
2. write – (communicate or express by writing; "Please write to me every week")
3. publish, write – (have (one's written work) issued for publication; "How many books did Georges Simenon write?"; "She published 25 books during her long career")
4. write, drop a line – (communicate (with) in writing; "Write her soon, please!")

Rather than being different senses, synsets 1 and 3 seem to be repetitions of exactly the same sense, associated with different synonyms. This is an example of a very common problem in WordNet. We refer the reader to Pustejovsky's argument (Pustejovsky 1995) against a "sense-enumerative lexicon" (one that enumerates different facets of the same sense as separate senses), and Wierzbicka's advice to lexicographers to "seek the invariant" (Wierzbicka 1993).

WordNet's synsets are structured hierarchically in an ontology. The nodes in this hierarchy do not represent semantic classes, nor do those classes fulfill particular slots in verb

[*] This work was supported by NIH grant 5R01-LM0006649-5 to Prof. Pustejovsky at Brandeis University. Direct all correspondences to jamesp@cs.brandeis.edu

argument structure. Examination of the superordinates (hyperonyms) of the ten synsets of *write* illustrate how WordNet fails to supply this sort of information:

(1) create verbally; (2) communicate, intercommunicate; (3) create verbally; (4) correspond; (5) create verbally; (6) make, create (which is itself a superordinate of "create verbally"); (7) trace, draw,line,describe,delineate; (8) record, tape; (9) [No superordinate]; (10) create by mental act, create mentally;

Even if the hierarchy of semantic types were to be pared down and reorganized – as they have been in EuroWordNet ((Vossen 1998)) – the nodes in the hierarchy, with their current populations of words, do not to show the associations between word senses and syntagmatic patterns, failing to identify the words needed to express the latter.

FrameNet

Fillmore's work in case grammar and frame semantics serves as a reminder of the holistic nature of verb argument structure, with alternations in the syntactic slots in which a particular semantic argument may be realized. FrameNet (Atkins, Fillmore, & Johnson (2003), Fillmore, Johnson, & Petruck (2003), Baker & Ruppenhofer (2002)) aims to translate those insights into a database of semantic frames, in which all the case roles implied by the semantics of each word are both stated and exemplified explicitly (regardless of whether they necessarily occur in all sentences in which the word is use).

FrameNet gives examples drawn from corpus data, but its analysis is not corpus-driven. It proceeds frame by frame, rather than word by word. It relies on the intuitions of its researchers to populate each frame with words. This runs the risk of accidental omissions and means that (in principle) no word can be regarded as completely analyzed until all frames are complete. At the time of writing, there has been no indication of when that will be, nor of the total number of frames that there will be. Currently, some frames overlap to the point of being indistinguishable (e.g., in the case of the verb *fire*). Others are only partly populated. Unfortunately, some frame announce a lexical entry as complete, when in fact only minor or rare senses have been covered.

FrameNet's methodology, which requires the researchers to think up all possible members of a Frame a priori, means that important senses of words that have been partly analysed are missing and may continue to be missing for years to come. There is no attempt in FrameNet to identify the senses of each word systematically and contrastively. In its present form, at least, FrameNet has at least as many gaps as senses. For example, at the time of writing *toast* is shown as part of the *Apply_Heat* frame but not in any of the frames that. include *applaud*, *praise*, or *celebration*. It is not clear how or whether the gaps are to be filled systematically. What is needed is a principled fix – a decision to proceed from evidence, not frames. This is ruled out by FrameNet for principled reasons: the unit of analysis for FrameNet is the frame, not the word.

VerbNet

In 2005, a first release of a new resource called VerbNet became available as described in Dang, Kipper, & Palmer (2000). VerbNet takes verb entries from Levin (1993) and adds thematic roles under the influence of Fillmore's Frame Semantics. Comparison of VerbNet entries with large-scale corpus evidence reveals some fundamental shortcomings in VerbNet. These include:

1. No clear definition of word sense Dang, Kipper, & Palmer (2000) acknowledge that polysemy is a "controversial area" in lexicon building and they say that they "address this problem by employing compositional semantics", but they do not show how this relates to actual usage.

2. Major senses missing The verb *fire* is in VerbNet only as a "throw verb" (Levin class 17.1), with the themantic roles "Destination[+animate], Source[+location] Theme[+concrete]", which is evidently intended as a representation of the Firearms sense. VerbNet does not cover other senses of *fire*, for example *dismiss*, as in *The company fired all its employees*.

3. Duplication of senses For example, VerbNet has three senses of *write*, which is a considerable improvement on the ten senses in WordNet cited above, but still two too many. There is no empirically satisfactory way of distinguishing between *write* as a *performance* (Levin class 26.7), *scribble* (Levin class 25.2), or *transfer-message* (Levin class 37.1)

4. Errors in thematic roles and other details In the entry for *fire*, "Destination[+animate]" is presumably intended to represent the target fired at. If so, this is a severe overrestriction – there are plenty of examples of locations and physical objects being fired at.

These are not isolated examples, but exemplify a deep-rooted and widespread problem throughout VerbNet, arising from reliance on sources that are not empirically well founded.

Generative Lexicon

Generative Lexicon (GL) is a theory of linguistic semantics which focuses on the distributed nature of compositionality in natural language (Pustejovsky, 1995). Unlike purely verb-based approaches to compositionality, it attempts to spread the semantic load across all constituents of an utterance. From the nature of word meaning to lexical creativity, GL provides a different perspective on many of NLP's most important questions. Hence, GL is not just a theory, but is meant to be implemented as a component of the backbone of larger NL systems (Pustejovsky and Boguraev, 1993). At the heart of GL is its network of qualia relations, and any true GL implementation would have to have a system of qualia-like structures. However, in current GL implementations people have found it difficult to integrate GL, with its large network of qualia relations, into large NL systems, since creating such an ontology requires a prohibitive investment of time and resources.

Brandeis Semantic Ontology

To help overcome this problem, we have been developing a large generative lexicon ontology and dictionary for use by the general research community. This system, called the

Brandeis Semantic Ontology (BSO), is intended to allow for more widespread access to GL-based lexical resources and help researchers in a variety of computational tasks. The specification of the type system used in the BSO largely follows that proposed by the SIMPLE specification (Busa et al., 2001), which was adopted by the EU-sponsored SIMPLE project (Lenci et al, 2000).

Following standard assumptions in GL, the computational resources available to a lexical item consist of four levels: Lexical Typing Structure; Argument Structure; Event Structure; and Qualia Structure. Qualia Structure is viewed as expressing the componential aspect of a word's meaning (Calzolari, 1992) and the meeting point of both argument and event structure. It is generally composed of the following attributes:

(1) a. FORMAL: the basic type distinguishing the meaning of a word;
b. CONSTITUTIVE: the relation between an object and its constituent parts;
c. TELIC: the purpose or function of the object, if there is one;
d. AGENTIVE: the factors involved in the object's origins or "coming into being".

The SIMPLE-GL model defines a language for making types, where qualia can be unified to create more complex concepts out of simple ones. Following Pustejovsky (2001), the ontology divides the domain of individuals into three levels of type structure:

(2) a. NATURAL TYPES: Natural kind concepts consisting of reference only to Formal and Const qualia roles;
b. FUNCTIONAL TYPES: Concepts making reference to purpose, function, or origin.
c. COMPLEX TYPES: Concepts integrating reference to a relation between types.

For example, a simple natural physical object (3), can be given a function (i.e., a Telic role), and transformed into a functional type, as in (4).

$$(3)\quad \begin{bmatrix} \textbf{physobj(x)} \\ \textsc{formal} = \textbf{physform(x)} \end{bmatrix}$$

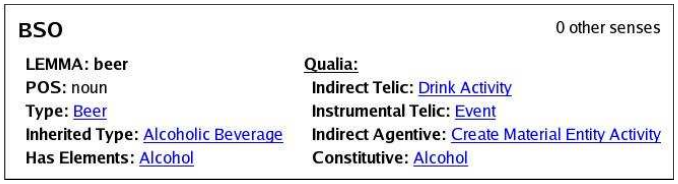
$$(4)\quad \begin{bmatrix} \textbf{artifact\_obj(x)} \\ \textsc{formal} = \textbf{physform(x)} \\ \textsc{telic} = \textbf{Pred(E,y,x)} \end{bmatrix}$$

Functional types (the "unified types" in Pustejovsky, 1995) behave differently from naturals, as they carry more information regarding their use and purpose. For example, the noun *sandwich* contains information of the "eating activity" as a constraint on its *Telic* value, due to its position in the type structure; that is, **eat(P,w,x)** denotes a process, **P**, between an individual **w** and the physical object **x**. It also reflects the fact that it is an artifact of a "making activity".

$$(5)\quad \begin{bmatrix} \textbf{sandwich(x)} \\ \textsc{const} = \{\textbf{bread,...}\} \\ \textsc{formal} = \textbf{physform(x)} \\ \textsc{telic} = \textbf{eat(P,w,x)} \\ \textsc{agentive} = \textbf{make\_activity(z,x)} \end{bmatrix}$$

Complex types, such as *book* and *university* are given a unique status in the BSO, implemented as product-types in order to capture the behavior of orthogonal inheritance (Pustejovsky and Boguraev, 1993). Examples of these types will be provided in the full paper.

To illustrate the nature of the semantic information in the BSO, consider the entry for the noun *beer*. The core features of this entry are shown below.



BSO 0 other senses

LEMMA: beer
POS: noun
Type: Beer
Inherited Type: Alcoholic Beverage
Has Elements: Alcohol

Qualia:
Indirect Telic: Drink Activity
Instrumental Telic: Event
Indirect Agentive: Create Material Entity Activity
Constitutive: Alcohol

In addition to its type, [[BEER]], and its inherited type, [[ALCOHOLIC_BEVERAGE]], it also displays the qualia associated with this type.

Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) is a technique that provides insight into the types of context parameters that allow humans to distinguish between different predicate senses (Pustejovsky, Hanks, & Rumshisky 2004). The CPA approach uses a semi-automatic bootstrapping process to produce selection contexts for predicates, extending the traditional notion of selectional context to include:

- shallow semantic typing of predicate arguments
- minor syntactic categories (locatives, adjuncts, etc.)
- predicate arguments represented by lexical sets
- subphrasal syntactic cues: genitives, partitives, bare plural/determiner, infinitivals, negatives, collocational cues

The procedure consists of three components: (1) the manual discovery of selection context patterns for specific verbs; (2) the automatic recognition of instances of the identified patterns; and (3) automatic acquisition of patterns for unanalyzed cases. During the lexical discovery stage, an analysis of corpus data is performed for a target lemma. The contexts of its usage are sorted into groups, and a stereotypical CPA pattern that captures the relevant semantic and syntactic features of the group is recorded. Each pattern is specified in terms of lexical sets for each argument, shallow semantic typing of these sets, and other syntagmatically relevant criteria, such as adverbials of manner, phrasal particles, genitives, negatives, etc. There is typically a many-to-one relation between the patterns and the senses they represent. The distribution of frequencies associated with each sense are typically far from even.

For example, here are some selected CPA patterns for the verb *fire*. *Fire* patterns representing senses that account for more than 5% of use are listed below.[1]

Selected CPA Patterns for FIRE:

```
I DISCHARGE A GUN AT A TARGET  (60%)

1. [[Person]] fire [[Artifact=Firearm]] (at [[PhysObj]])

2. [[Person]] fire [[Artifact=Projectile]] (off)
   ({from [[Artifact=Firearm]]}) ({at [[PhysObj]]}
   | [ADV[Direction]])

3. [[Person]] fire [NO OBJ] ({at [[PhysObj]]}
   | {on [[HumanGroup]]} | [ADV[Direction]])

4. [[Artifact=Firearm]] fire [NO OBJ] ({at [[PhysObj]]}
   | {on [[HumanGroup]]} | [Adv[Direction]])

III DISMISS AN EMPLOYEE (11%)

6. [[Person 1]] fire [[Person 2]] (for [[Action=Bad]])

VII INSPIRE SOMEONE (11%)

12. [[TopType]] fire {[[Person]]'s [[Attitude=Enthusiasm]]}

13. [[TopType]] fire [[Person]] (up)
```

The CPA approach has its origins in the analysis of large corpora for lexicographic purposes (e.g. Cobuild (Sinclair, Hanks, & al. 1987)). Its objective is to identify, in relation to a given target word, the overt textual clues that activate one or more components of its meaning potential. CPA is concerned with establishing prototypical norms of usage for individual words and it offers a contrastive analysis of the senses of each word.

CPA Sense Differentiation

Any number of context dimensions can affect semantic interpretation of a lemma. For example, the presence or absence of an argument such as adverbial of manner, selects different meanings of the verb *treat* (Pustejovsky & Hanks 2001). `[[Person 1]] treat [[Person 2]]` without an adverbial of manner generally activates the medical sense of *treat*, while `[[Person 1]] treat [[Person 2]] [Adv[Manner]]` activates the behavior-attitude sense of this verb (e.g., *She treated him with respect*).

The most frequent source of meaning differentiation of verbs lies in contrasting the argument types filling each argument slot. Whenever a semantic type based specification of predicate arguments is insufficient for the purposes of sense distinction, predicate arguments have to be specified in terms of lexical sets, i.e., by enumerating typical members.

Thus, *gun, rifle, MK17, Kalashnikov, pistol, revolver, cannon* are canonical members of a lexical set [[Firearm]]. These words form a paradigmatic cluster, i.e., they tend to co-occur in the same argument slot of different verbs. This cluster also corresponds to the type [[Firearm]] in the Brandeis Semantic Ontology. Note that the same lexical set also contains outliers, e.g., *bow, catapult* which fulfill the same semantic role as *gun* in relation to the verb *fire*, even though they are not, strictly speaking, firearms.

Semantic restrictions[2] imposed by verbs on their arguments vary greatly in size and scope. In some cases, there could be few or no lexical preferences and the semantic value of the argument is expressed as `[[TopType]]`, as is the case with subject in the pattern `[[TopType]] fire [[Person]]'s [[Attitude=Enthusiasm]]`. This essentially means that almost anything can (and regularly does) fire a person's enthusiasm. The direct object slot for this pattern consists of a much narrower lexical set, where the canonical members are given by *enthusiasm, imagination, interest.*

Integration of CPA and BSO

In this section we come finally to the merger of the CPA technique with the representations as they exist in the BSO. As stated earlier, our goal is to contextualize the semantic information associated with a type and the words associated with that type (through information derived from corpus analysis.) In particular, we will illustrate how the arguments of predicates are given type specifications that correspond to semantic tags from CPA output. Secondly, we illustrate briefly how the lexical extensions of the entity types in the ontology are modified and verified through inclusion in the lexical sets from the CPA.

The integration proceeds in two steps and two directions: (1) mapping CPA patterns to BSO verb-argument entries; and (2) mapping BSO entity type extensions to the lexical sets in CPA. In the discussion that follows, we describe each of these steps.

The BSO contains type assignments for 20,000 noun entries and 10,000 nominal collocation entries[3]. Currently, a projection of BSO, called BSO Lite, is used by CPA to help identify the lexical sets of predicate arguments with particular semantic types. The BSO Lite is a shallow hierarchy of types selected for their prevalence in manually identified selection context patterns and their utility in discriminating predicate senses in the corpus. At the time of writing, there are 65 types which contribute a basic structure in terms of which patterns for the first hundred verbs have been analyzed.

The BSO Lite has been used with CPA-derived context features to improve disambiguation of a subset of Senseval verbs (Rumshisky & Pustejovsky 2006). However, in the contextualized version of the BSO, the arguments of these verbs are linked to the appropriate nodes in the entity hierarchy, thereby allowing for direct disambiguation.

[1] See (Pustejovsky, Hanks, & Rumshisky 2004) for pattern syntax.

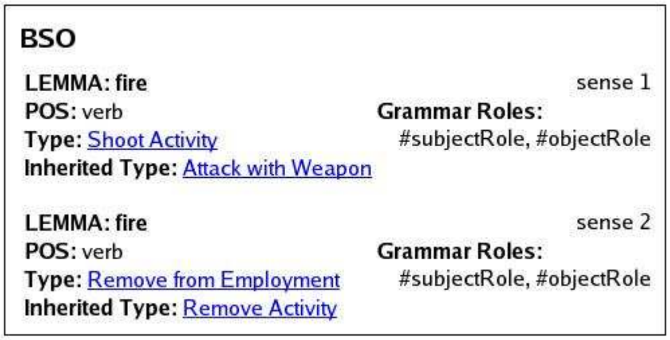
[2] All semantic restrictions are of course probabilistic.

[3] The BSO ontology was compiled as an internal lexical database on the basis of pre-corpus resources, supported by commonsense intuitions. Many of the revisions now being made to BSO in the light of CPA's empirical data involve complex restructuring.

The organization of nouns in the BSO follows GL principles of linking the type of an entity to relational types, e.g., the qualia structure. For example, a noun such as *gun* is typed as a weapon, and as such inherits the *Telic* quale associated with the use of a weapon in firing, attacking, etc. That is, the entities point directly to relations they are involved in prototypically or conventionally. Hence, what we have is a normal type lattice along with the additional links provided by the qualia.

The organizing principle for the verb hierarchy in the BSO provides for identification of the arguments to a predicate, but does not link these arguments explicitly to nodes in the entity hierarchy. Our first objective in merging these approaches is to enrich the verb argument specification with the semantic typing information from the CPA context patterns. As a result of this technique, we are also able to identify the appropriate level of generalization in the type system for noun classes, as a result of knowing what semantic features are relevant for distinguishing the senses of a predicate. This is, in fact, what the contextualization of a predicate relative to a corpus is.

For example, consider the semantics of the verb entry *fire* in the BSO. There are two main verb senses in the BSO for *fire*:



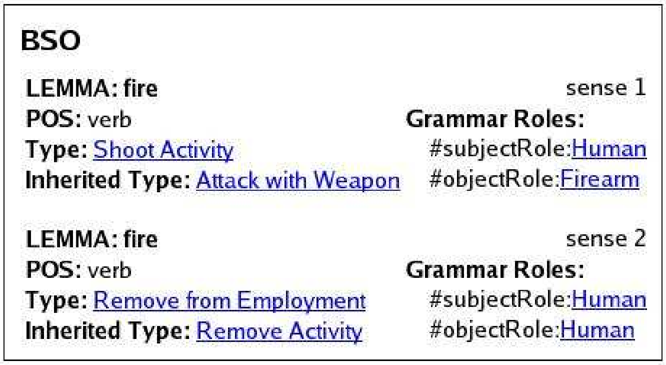
BSO

LEMMA: fire sense 1
POS: verb
Type: Shoot Activity
Inherited Type: Attack with Weapon
Grammar Roles:
#subjectRole, #objectRole

LEMMA: fire sense 2
POS: verb
Type: Remove from Employment
Inherited Type: Remove Activity
Grammar Roles:
#subjectRole, #objectRole

As stated above, while the argument structure associated with each sense of *fire* is specified, no explicit semantic typing on these arguments is assigned. The two senses of the verb are distinguished by their local typing. Our aim here is to unpack the verb meaning from its type into a fully specified argument structure with selectional typing assigned to the arguments. That is, the verb has two senses by virtue of its argument specification.

To illustrate this process, consider the result of contextualizing the entry for the verb *fire* in the BSO is shown below:



BSO

LEMMA: fire sense 1
POS: verb
Type: Shoot Activity
Inherited Type: Attack with Weapon
Grammar Roles:
#subjectRole:Human
#objectRole:Firearm

LEMMA: fire sense 2
POS: verb
Type: Remove from Employment
Inherited Type: Remove Activity
Grammar Roles:
#subjectRole:Human
#objectRole:Human

The two senses which are distinguishable by name of the semantic type have been unpacked through this process, and the implicating features have been anchored to the appropriate arguments to the verb.

Approximately 900 patterns have been identified for over a hundred verbs through the CPA technique. The above contextualization procedure is currently being applied to the corresponding verb forms in the BSO. As a result of this procedure, the senses that exist in the BSO will be modified, deleted or supplemented according to the empirical findings.

A similar verification and modification procedure is applied to the [[ENTITY]] types in the BSO. Nouns denoting [[ENTITIES]] are grouped together and typed according to their tendency to co-occur in the same argument slot in relation to verbs. The goal is to substantiate the existing type hierarchy and restructure it where necessary by verifying the lexical extensions for each type. A semantic type is retained in the ontology if it carries a semantic feature that verifiably contributes to actual sense distinctions observed in the corpora through CPA analysis.

During the verification procedure, lexical sets that are organized on the basis of a particular semantic feature (rather than, say, an additional value or a role assigned to the argument) are linked to or associated with a specific BSO type, thereby verifying its utility and validity of its extension.

For example, consider the list of artifacts whose *Telic* is *to be fired* and which therefore count as [[FIREARMS]]. The original BSO contains the following set of lexical items for the semantic type [[FIREARM]]: *38 caliber, 22 rifle, 38 caliber, ack-ack gun, air gun, antiaircraft gun, fire ship, gas gun, greek fire, minute gun, quaker gun, set gun, spring gun, whaling gun, ack-ack, airgun, antiaircraft, arquebus, automatic, cannon, colt, firearm, flak ,flamethrower, gun, handgun, pistol, revolver, rifle, six-gun, six-shooter, small-arm*. The following lexical items are removed from this set on the grounds that their syntagmatic behavior is different from that of other members of the set: *fireship, antiaircraft, flak, small arm*. *Flak*, for example, belong in the [[PROJECTILE]] set, not the [[FIREARM]] set. At the same time, *howitzer, mortar, machine gun, submachine gun, Bren gun, Kalashnikov*, and several other items are added on the grounds that these are canonical members of the lexical set of [[FIREARMS]].

Note that some of the lexical items that belong to the BSO type [[BOW]] (*bow, crossbow, longbow*, etc.), the sibling of type [[FIREARM]], are in fact canonical members of the lexical set [[FIREARM]]. Although the technology involved in firing a bow is different from the technology of firing a gun, the linguistic behavior of both sets is very similar.

It should be pointed out that semantic restrictions placed by predicates on their arguments vary not just in type specificity (e.g., [[FIREARM]] vs. [[ARTIFACT]]), but also in properties extending over disjoint types. For example, in interpreting the lexical sets for the arguments to the verb *risk*,

(1) a. risk one's *life/money/name/reputation/...*
b. risk *death/bankruptcy/injury/wrath/backlash/...*

the relevant semantic restrictions placed on the direct object is not the type [[EVENT]] alone, but rather the property reflecting the negative value judgment associated with it (i.e., *Bad*)[4].

This aspect of CPA methodology highlights that conceptual connections may exist between entities of various types that are not representable in a conventional type hierarchy, nor, as it happens, in the existing data structures of GL. For example, values of *Good* and *Bad* are not currently represented in the qualia structure in the BSO. But clearly, there is a cognitive salience to such categorization as reflected in the empirical data.

Contextualizing ontologies to novel corpora CPA analysis is a corpus-based technique that allows one to contextualize an ontology to any domain. For example, this technique is easily extended into the domain of biomedical literature. In fact, the work we present here extends the work we began with ontology rerendering, which involved adapting ontology to a corpus of biomedical literature (Pustejovsky, Rumshisky, & Castano 2002).

Rerendering procedure used two strategies, namely, noun phrase-based rerendering and biorelation-based rerendering. In both cases, original corpus is tagged with the seed ontology. In the first strategy, extensions to the nodes of the original ontology are created through frequency analysis of noun phrase structure. In the second strategy, ad-hoc categories are created through a statistically thresholded, typed projection of the arguments of biorelation. The resulting adhoc category is then matched with the types obtained at the second-level of NP-based ontology extension.

The second step in the rerendering procedure links adhoc categories of arguments in the same way predicate types are contextualized to corpus-derived argument sets that CPA provides.

In the biological literature, verbs of biological interaction, such as *inhibit* or *activate* are semantically underspecified, and a lexicon for such predicates is uninformative with respect to the specificity of the interaction. The specific biological interactions come only through the details of the actual arguments participating in the interaction. Such constraints are captured by CPA through the membership in lexical sets for the arguments.

[4] This, in fact, is a reflection from corpus data of the coercion analyzed for such verbs in (Pustejovsky 1995).

Conclusion

In this paper, we demonstrated how existing lexical resources for linguistic knowledge fall short of modeling the behavior of lexemes in actual text corpora. To remedy this, we presented a methodology for contextualizing a linguistically designed lexical resource (Brandeis Semantic Ontology) through the integration of of a corpus-based lexical acquisition process, called Corpus Pattern Analysis. This results in an ontology of types for lexical items that is more attuned to the semantic selectional environment for a specific word sense. We hope to continue the application of this methodology over the entire type system of the BSO.

References

Atkins, S.; Fillmore, C.; and Johnson, C. 2003. Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography* 16(3):251–280.

Baker, C., and Ruppenhofer, J. 2002. Framenet's frames vs Levin's verb classes. *28th Annual Meeting of the Berkeley Linguistic Society*.

Busa, F.; Calzolari, N.; and Lenci, A. 2001. Generative lexicon and the SIMPLE model: Developing semantic resources for nlp. In *The Syntax of Word Meaning*. Cambridge University Press.

Dang, H.; Kipper, K.; and Palmer, M. 2000. Integrating compositional semantics into a verb lexicon. In *Seventeenth National Conference on Artificial Intelligence AAAI 2000*.

Fillmore, C.; Johnson, C.; and Petruck, M. 2003. Background to Framenet. *International Journal of Lexicography* 16(3):235–250.

Hanks, P., and Pustejovsky, J. 2005. A pattern dictionary for natural language processing. *Revue Fran caise de Linguistique Appliquée*.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Pustejovsky, J., and Hanks, P. 2001. Very Large Lexical Databases: A tutorial. *ACL Workshop, Toulouse, France*.

Pustejovsky, J.; Havasi, C.; Saurí, R.; Hanks, P.; and Rumshisky, A. 2005. Towards a generative lexical resource:The Brandeis Semantic Ontology. *Submitted to LREC 2006, Genoa, Italy*.

Pustejovsky, J.; Hanks, P.; and Rumshisky, A. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, 924–931.

Pustejovsky, J.; Rumshisky, A.; and Castano, J. 2002. Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics. In *LREC 2002 Workshop on Ontologies and Lexical Knowledge Bases. Las Palmas, Canary Islands, Spain*.

Pustejovsky, J. 1995. Generative Lexicon. *Cambridge (Mass.): MIT Press*.

Rumshisky, A., and Pustejovsky, J. 2006. Inducing sense-discriminating context patterns from sense-tagged corpora. In *LREC 2006, Genoa, Italy*.

Sinclair, J.; Hanks, P.; and al. 1987. *The Collins Cobuild English Language Dictionary*. HarperCollins, 4th edition. Collins Cobuild Advanced Learner's English Dictionary (2003).

Vossen, P. 1998. Introduction to EuroWordNet. *Computers and the Humanities* 32(2-3):73–89.

Wierzbicka, A. 1993. What are the uses of theoretical lexicography? *Dictionaries: Journal of the Dictionary Society of North America* 14:51–57.