Towards a Generative Lexical Resource:
The Brandeis Semantic Ontology

James Pustejovsky, Catherine Havasi, Jessica Littman,
Anna Rumshisky, Marc Verhagen

Computer Science Dept., Brandeis University
415 South St.
Waltham, MA, 02454
{jamesp, havasi, roser, arum, jlittman, marc}@cs.brandeis.edu

Abstract

In this paper we describe the structure and development of the Brandeis Semantic Ontology (BSO), a large generative lexicon ontology and lexical database. The BSO has been designed to allow for more widespread access to Generative Lexicon-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).

1. Introduction

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, GL 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 NLPs 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). One of the major complaints against GL has been that it is too difficult to embed in such a system. 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, creating such an ontology requires a prohibitive investment of time for most researchers, and current GL implementations place large constraints on any lexical development project.

To help overcome this problem, we have developed 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).

2. Lexical Design

The ontology consists of three major types: entity, event, and property. Here, we focus on the entity and event hierarchies. Each of these is divided into three further hierarchies: natural, artifactual, and complex. This type classification is described in (1).

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

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:

(2) 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".

For purposes of database design, type assignment is represented as predication of that type, be it natural, artifactual, or complex. For example, a simple natural physical object (3), can be given a function (i.e., a Telic role), and transformed into an artifactual type, as in (4).

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

$$(4) \quad \begin{bmatrix} \textbf{artifact\_obj(x)} \\ \text{FORMAL} = \textbf{physform(x)} \\ \text{TELIC} = \textbf{Pred(E,y,x)} \end{bmatrix}$$

Artifactual 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 that it is an artifact of a "making activity".

$$(5) \quad \begin{bmatrix} \textbf{sandwich(x)} \\ \text{CONST} = \{\textbf{bread,...}\} \\ \text{FORMAL} = \textbf{physform(x)} \\ \text{TELIC} = \textbf{eat(P,w,x)} \\ \text{AGENTIVE} = \textbf{make\_activity(z,x)} \end{bmatrix}$$

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). Complex types require the full qualia structure from both of their formal types. In this way they are different from artifactual types that simply map back to a natural type (see below). Example (6) shows the qualia structure for *book*.

$$(6) \quad \begin{bmatrix} \textbf{book(x)} \\ \text{FORMAL} = \textbf{physform} \bullet \textbf{information(x)} \\ \text{TELIC} = \textbf{read(P,w,x)} \\ \text{AGENTIVE} = \textbf{write(z,x)} \end{bmatrix}$$

Since the entity and event types in the BSO are divided into three sub-hierarchies, we need to account for inheritance both within a hierarchy and from one to the other. Within the naturals, inheritance runs along formal lines. In the artifactual hierarchy, inheritance runs through the telic quale. We refer to this kind of inheritance as "subtyping". An entry in the artifactual hierarchy, such as *doctor*, maps back to the natural hierarchy through its formal quale, *human*. This is called "formal mapping". Entries in the complex hierarchy can formal map back to several different entries in the other hierarchies either by way of a telic quale to the artifactual hierarchy or through a formal quale to the naturals. The development of the BSO will be tied to the Corpus Pattern Analysis project (Rumshisky et al., 2006). This project aims to find patterns that reflect normal usage of language. CPA and BSO are in simultaneous development and will influence each other in the following ways:

1. CPA patterns refer to types in the BSO

2. CPA patterns suggest changes to the BSO

3. CPA is a major driving force in the creation of the BSO event hierarchy

As the CPA project progressed, a list of shallow types has been formed. These shallow types make up the most shallow levels of the natural and artifactual hierarchies of the BSO. In addition to a shallow type, an entry in CPA also includes a subspecification that helps distinguish among senses. The example in (7) shows a typical CPA entry.

(7) [[person=doctor]] "treats" [[person=patient]]

For each argument of *treat* a semantic type from the shallow ontology is given and a subspecification that can either be an implicature, a lexical set, or some property. This CPA pattern will relate to BSO entries for *doctor* and *patient* as well as for the event *treat*. The semantic type that is given for each argument will be the general type in the BSO entry while the subspecification can be a type from anywhere in the three hierarchies as long as it is connected to the general type through subtyping or formal mapping. The following scenarios describe some ways in which CPA informs the BSO:

Scenario 1 A lexicographer analyzes the corpus patterns for *drink* and sees the sentence: *he drank beer*. He clicks *beer* and is presented with a list from the BSO with all types and supertypes for *beer*. For each type, a list of entries can be displayed. He finds the type [[Beverage]] and decides that that type, with its entries, captures a normal use of *drink*, and then the following pattern can be created:

drink_2:[[Person]] drinks [[Beverage]]

The BSO is automatically notified of this action, possibly by adding two entries to a table associated with the BSO types:

type	pattern	role
Person	drink_2	subject
Beverage	drink_2	object

These entries suggest to a BSO editor that [[Person]] and [[Beverage]] are useful types that should not be changed lightly. Furthermore, CPA should be consulted or warned if these BSO types are changed.

Scenario 2 A lexicographer analyzes the corpus patterns for *drink* and sees the sentence *he drank beer*. She clicks *beer* and is presented with the appropriate BSO list. She finds the [[Beverage]] and decides that the type is fine, but that only a subset of its lexical entries is relevant. For example, she believes that *beer* and *ale* are appropriate while *ginger ale* and *lager* are not good for this pattern, though they are still good beverages and useful for other patterns. The lexicographer proposes the following pattern:

drink_3:[[Person]] drinks [[Beverage]]{"beer","ale"}

BSO is notified of this action, possibly by the addition of the following entries in a table associated with the BSO entries:

entry	type	pattern	role
beer	Beverage	drink_3	object
ale	Beverage	drink_3	object

These entries suggest to a BSO editor that *beer* and *ale* of type [[Beverage]] are useful entries with a useful a type and they should not be changed lightly. CPA should be consulted or warned when the BSO types are changed.

In the following examples, we see how some entries in the BSO look.

(8) [[drink activity]]
    supertype = [[Take Nourishment Activity]]
    #subject = [[Animate Living Entity]]
    #object = [[Beverage]]

(9) 'drink'
    type = [[Drink Activity]]

Some entries will set restrictions on types of roles. For example, *chug* is a particular kind of drinking and selects for a particular kind of beverage:

(10) 'chug'
    type = [[Drink Activity]]
    #object = [[Alcoholic Beverage]]

The above examples are taken from the event types. In examples (11-13) we see how some entries of type entity may look.

(11) [[Writer]]
    #telic = [[Write Activity]]

(12) [[Write Activity]]
    #object = [[Book]]

(13) 'novelist'
    type = [[Writer]]
    (#telic -> #object) = [[Novel]]

Although the BSO is based in large measure on the SIMPLE-GL specification, it differs from other previously developed GL-lexicons in some important respects. The SIMPLE project was a EU-sponsored effort to annotate GL inspired information in parallel lexicons for the various European languages. However, since the English SIMPLE lexicon has never been completed, researchers have been left without a SIMPLE for the language of many of the popular corpora. The BSO is focused on the English lexicon alone, and the current BSO size is larger than any of the SIMPLE implemented lexicons. EuroWordNet (Vossen, 1998), a project to create semantic networks for the major European languages, is influenced in part by GL, but in contrast to the BSO, it does not include qualia for its words.

3. Release Schedule

By May, 2006 we will have the BSO ready for a public release, at which point it will be accessible to researchers in computational linguistics and other related fields, either as a centrally located database or as a downloadable package. By then we hope to have 30,000 qualia-related lexical entries and 4,500 entries in the ontological type system. After the release, we aim to continue to improve this lexical resource by adding additional lexical information and improving the type system. Along with the database, we will

release a series of browsers as well. We have created both a stand-alone browser, BULB, as well as an online browsing tool (Havasi et al., 2006). Both browsers display the lexical information about a selected word in the BSO as well as the word's position in the ontological type system. Additional functionality provides comparisons to other lexical resources, such as WordNet, FrameNet, and PropBank. We

believe that this system will be of use in the various commonly studied problems in NLP, as well as in areas yet to be explored. We feel this work would be useful to the task of textual entailment, as determining the relationship between two sentences would be aided by further semantic understanding of the relationships between words in them. Question answering and summarization tasks would similarly be aided by a better understanding of the structure of the lexical relations between the words with which they are dealing. This could also be useful to the word sense disambiguation community by providing richer lexical context for words. Finally, the information retrieval community has expressed interest in GL, but wishes for "more than a formalism" (Claveau et al., 2003) which they could work with, and the BSO could provide the concrete implementation they are looking for.

4. References

F. Busa, N. Calzolari, and A. Lenci, 2001. *Generative Lexicon and the SIMPLE Mode: Developing Semantic Resources for NLP*. Cambridge University Press, Cambridge.

N Calzolari, 1992. *Acquiring and Representing Semantic Information in a Lexical Knowledge Base*. Springer Verlag, New York.

V. Claveau, P. Sébillot, C. Fabre, and P. Bouillon. 2003. Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming. *Journal of Machine Learning Research*, 4:493–525.

Catherine Havasi, James Pustejovsky, and Marc Verhagen. 2006. Bulb: A unified lexical browser. In *Proceedings of LREC 2006*, Genoa, Italy.

A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. 2000. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

James Pustejovsky and Bran Boguraev. 1993. Lexical knowledge representation and natural language processing. *Artificial Intelligence*, 63:193–223.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Anna Rumshisky, Patrick Hanks, Catherine Havasi, and James Pustejovsky. 2006. Constructing a corpus-based ontology using model bias. In *Proceedings of FLAIRS*, Melbourne, FL.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands.