The Acquisition of Lexical Semantic Knowledge from Large Corpora

James Pustejovsky

Computer Science Department
Brandeis University
Waltham, MA 02254

ABSTRACT

Machine-readable dictionaries provide the raw material from which to construct computationally useful representations of the generic vocabulary contained within it. Many sublanguages, however, are poorly represented in on-line dictionaries, if represented at all. Vocabularies geared to specialized domains are necessary for many applications, such as text categorization and information retrieval. In this paper I describe research devoted to developing techniques for building sublanguage lexicons via syntactic and statistical corpus analysis coupled with analytic techniques based on the tenets of a generative lexicon.

1. Introduction

Machine-readable dictionaries provide the raw material from which to construct computationally useful representations of the generic vocabulary contained within it. Many sublanguages, however, are poorly represented in on-line dictionaries, if represented at all (cf. Grishman et al (1986)). Yet vocabularies geared to specialized domains are necessary for many applications, such as text categorization and information retrieval. In this paper I describe research devoted to developing techniques for building sublanguage lexicons via syntactic and statistical corpus analysis coupled with analytic techniques based on the tenets of a generative theory of the lexicon (Pustejovsky 1991).

Unlike with purely statistical collocational analyses, the framework of a lexical semantic theory allows the automatic construction of predictions about deeper semantic relationships among words appearing in collocational systems. I illustrate the approach for the acquisition of lexical information for several lexical classes, and how such techniques can fine tune the lexical structures acquired from an initial seeding of a machine-readable dictionary, i.e. the machine-tractable version of the LDOCE (Wilks et al (1991)).

The aim of our research is to discover what kinds of knowledge can be reliably acquired through the use of these methods, exploiting, as they do, general linguistic knowledge rather than domain knowledge. In this respect, our program is similar to Zernik (1989) and Zernik and Jacobs (1990), working on extracting verb semantics from corpora using lexical categories. Our research, however, differs in two respects: first, we employ a more expressive lexical semantics; secondly, our focus is on all major categories in the language, and not just verbs. This is important since for full-text information retrieval, information about nominals is paramount, as most queries tend to be expressed as conjunctions of nouns. From a theoretical perspective, I believe that the contribution of the lexical semantics of nominals to the overall structure of the lexicon has been somewhat neglected, relative to that of verbs (cf. Pustejovsky and Anick (1988), Boguraev and Pustejovsky (1990)). Therefore, where others present ambiguity and metonymy as a potential obstacle to effective corpus analysis, we believe that the existence of motivated metonymic structures actually provides valuable clues for semantic analysis of nouns in a corpus. To demonstrate these points, I describe experiments performed within the DIDEROT Tipster Extraction project (of Brandeis University and New Mexico State University), over a corpus of joint venture articles.

2. Projecting Syntactic Behavior from Deep Semantic Types

The purpose of the research is to experiment with automatic acquisition of semantic tags for words in a sublanguage, tags which are well beyond that available from the seeding of MRDs. The identification of semantic tags for a word associated with particular lexical forms (i.e. semantic collocations) can be represented as that part of the lexical structure of a word called the projective conclusion space (Pustejovsky (1991))).

For this work, we will need to define several semantic notions. These include: *type coercion*, where a lexical item requires a specific type specification for its argument, and the argument is able to change type accordingly —this explains the behavior of logical metonymy and the syntactic variation seen in complements to verbs and nominals; *cospecification*, a semantic tagging of what collocational patterns the lexical item may enter into; and *contextual opacity/transparency*, which characterizes of a word just how it is used in particular contexts. Formally, we will identify this property with specific cospecification values for the lexical item (cf. Pustejovsky (forthcoming)).

Metonymy, in this view, can be seen as a case of the "licensed violation" of selectional restrictions. For example, while the verb *announce* selects for a human subject, sentences like *The Dow Corporation announced third quarter losses* are not only an acceptable paraphrase of the selectionally correct form *Mr. Dow Jr. announced third quarter losses for Dow Corp*, but they are the preferred form in the corpora being examined (i.e. the ACL-DCI WSJ and TIPSTER Corpora). This is an example of subject *type coercion*, where the semantics for Dow Corp. as a company must specify that there is a human typically associated with such official pronouncements (Bergler (forthcoming)).

2.1. Coercive Environments in Corpora

Another example of type coercion is that seen in the complements of verbs such as *begin*, *enjoy*, *finish*, etc. That is, in sentences such as "John began the book", the normal complement expected is an action or event of some sort, most often expressed by a gerundive or infinitival phrase: "John began reading the book", "John began to read the book". In Pustejovsky (1991) it is argued that in such cases, the verb need not have multiple subcategorizations, but only one *deep semantic type*, in this case, an event. Thus, the verb 'coerces' its complement (e.g. "the book") into an event related to that object. Such information can be represented by means of a representational schema called *qualia structure*, which, among other things, specifies the relations associated with objects.

In related work being carried out with Mats Rooth of ATT, we are exploring what the range of coercion types is, and what environments they may appear in, as discovered in corpora. Some of our initial data suggest that the hypothesis of deep semantic selection may in fact be correct, as well as indicating what the nature of the coercion rules may be. Using techniques described in Church and Hindle (1990), Church and Hanks (1990), and Hindle and Rooth (1991), below are some examples of the most frequent V-O pairs from the AP corpus.

```
Counts for "objects" of begin/V:
205 begin/V career/O
176 begin/V day/O
159 begin/V work/O
140 begin/V talk/O
120 begin/V campaign/O
113 begin/V investigation/O
106 begin/V process/O
 92 begin/V program/O
 85 begin/V operation/O
 85 begin/V negotiation/O
 66 begin/V strike/O
 64 begin/V production/O
 59 begin/V meeting/O
 59 begin/V term/O
 50 begin/V visit/O
 45 begin/V test/O
 39 begin/V construction/O
 31 begin/V debate/O
 29 begin/V trial/O
```

Corpus studies confirm similar results for "weakly intensional contexts" (Pustejovsky (1991)) such as the complement of coercive verbs such as *veto*. These are interesting because regardless of the noun type appearing as complement, it is embedded within a semantic interpretation of "the proposal to", thereby clothing the complement within an intensional context. The examples below with the verb *veto* indicate two things: first, that such coercions are regular and pervasive in corpora; secondly, that almost anything can be vetoed, but that the most frequently occurring objects are closest to the type selected by the verb.

```
303 veto/V bill/O
 84 veto/V legislation/O
 58 veto/V measure/O
 35 veto/V resolution/O
 21 veto/V law/O
 14 veto/V item/O
 12 veto/V decision/O
  9 veto/V proposal/O
  9 veto/V plan/O
  7 veto/V package/O
  6 veto/V increase/O
  5 veto/V sanction/O
  5 veto/V penalty/O
  4 veto/V notice/O
  4 veto/V idea/O
  4 veto/V appropriation/O
  4 veto/V mission/O
  4 veto/V attempt/O
  3 veto/V search/O
  3 veto/V cut/O
  3 veto/V deal/O
  1 veto/V expedition/O
```

What these data show is that the highest count complement types match the type required by the verb; namely, that one vetoes a bill or proposal to do something, not the thing itself. These nouns can therefore be used with some predictive certainty for inducing the semantic type in coercive environments such as "veto the expedition." This work is still preliminary, however, and requires further examination (Pustejovsky and Rooth (in preparation)).

3. Implications for Natural Language Processing

The framework proposed here is attractive for NLP, for at least two reasons. First, it can be formalized, and thus make the basis for a computational procedure for word interpretation in context. Second, it does not require the notion of exhaustive enumeration of all the different ways in which a word can behave, in particular in collocations with other words. Consequently, the framework can naturally cope with the 'creative' use of language; that is, the open-ended nature of word combinations and their associated meanings.

The method of fine-grained characterization of lexical entries, as proposed here, effectively allows us to conflate different word senses (in the traditional meaning of this term) into a single *meta-entry*, thereby offering great potential not only for systematically encoding regularities of word behavior dependent on context, but also for greatly reducing the size of the lexicon. Following Pustejovsky and Anick (1988), we call such meta-entries *lexical conceptual paradigms (LCPs)*. The theoretical claim here is that such a characterization constrains what a possible word meaning can be, through the mechanism of logically well-formed semantic expressions. The expressive power of a KR formalism can then be viewed as simply a tool which gives substance to this claim.

The notion of a meta-entry turns out to be very useful for capturing the systematic ambiguities which are so pervasive throughout language. Among the alternations captured by LCPs are the following (see Pustejovsky (forthcoming) and Levin (1989)):

1. Count/Mass alternations; e.g. *sheep*.
2. Container/Containee alternations; e.g. *bottle*.
3. Figure/Ground Reversals; e.g. *door*, *window*.
4. Product/Producer diathesis; e.g. *newspaper*, *IBM*, *Ford*.
5. Plant/Food alternations; e.g. *fig*, *apple*.
6. Process/Result diathesis; e.g. *examination*, *combination*.
7. Place/People diathesis; e.g. *city*, *New York*.

For example, an apparently unambiguous noun such as *newspaper* can appear in many semantically distinct contexts.

1. *The coffee cup is on top of the newspaper.*
2. *The article is in the newspaper.*
3. *The newspaper attacked the senator from Massachusetts.*
4. *The newspaper is hoping to fire its editor next month.*

This noun falls into a particular specialization of the *Product/Produce*r paradigm, where the noun can logically denote either the organization or the product produced by the organization. This is another example of logical polysemy and is represented in the lexical structure for *newspaper* explicitly (Pustejovsky (1991)).

Another class of logically polysemous nominals is a specialization of the process/result nominalssuch as *merger*, *joint venture*, *consolidation*, etc. Examples of how these nominals pattern syntactically in text are given below:

1. *Trustcorp Inc. will become Society Bank & Trust when its merger is completed with Society Corp. of Cleveland, the bank said.*
2. *Shareholders must approve the merger at general meetings of the two companies in late November.*
3. *But Mr. Rey brought about a merger in the next few years between the country's major producers.*
4. *A pharmaceutical joint venture of Johnson & Johnson and Merck agreed in principle to buy the U.S. over-the-counter drug business of ICI Americas for over $450 million.*
5. *The four-year-old business is turning a small profit and the entrepreneurs are about to sign a joint venture agreement with a Moscow cooperative to export the yarn to the Soviet Union.*

Because of their semantic type, these nominals enter into an LCP which generates a set of structural templates predicted for that noun in the language. For example, the LCP in this case is the **union** concept, and has the following lexical structure associated with it:

```
LCP: type: union
 [ Const: >2x:entity(x) ]
 [ Form:  exist(1y) [entity(y)] ]
 [ Agent: type:event & join(x) ]
 [ Telic: nil ]
```

This states that a **union** is an event which brings about one entity from two or more, and comes about by a joining event. The lexical structure for the nominal *merger* is inherited from this paradigm.

```
merger(*x*)
  [ Const: ({w}>2) [company(w) or firm(w)] ]
  [ Form:  exists(y) [company(y)] ]
  [ Agent: event(*x*): join(*x*,{w}) ]
  [ Telic: contextual]
```

It is interesting to note that all synonyms for this word (or, alternatively, viewed as clustered under this concept) will share in the same LCP behavior: e.g. *merging*, *unification*, *coalition*, *combination*, *consolidation*, etc.

With this LCP there are associated syntactic realization patterns for how the word and its arguments are realized in text. Such a paradigm is a very generic, domain independent set of schemas, which is a significant point for multi-domain and multi-task NLP applications.

For the particular LCP of union, the syntactic schemas include the following:

LCP schemas:
[where N=UNION; X=arg1; Y=arg2]

N of X and Y
X's N with Y
Y's N with X
N between X and Y
N of Z (Z=X+Y)
N between Z

EXAMPLE:

merger of x and y
x's merger with y
y's merger with x
merger between x and y
merger of the two companies
merger between two companies

There are several things to note here. First, such paradigmatic behavior is extremely regular for nouns in a language, and as a result, the members of such paradigms can be found using knowledge acquisition techniques from large corpora (cf. Anick and Pustejovsky (1990) for one such algorithm). Secondly, because these are very common nominal patterns for nouns such as *merger*, it is significant when the noun appears without all arguments explicitly expressed. For example, in (5) below, presuppositions from the lexical structure combine with discourse clues in the form of definite reference in the noun phrase (*the merger*) to suggest that the other partner in the merger was mentioned previously in the text.

5. *Florida National said yesterday that it remains committed to the merger.*

Similarly powerful inferences can be made from an indefinite nominal when introduced into the discourse as in (6). Here, there is a strong presupposition that both partners in the merger are mentioned someplace in the immediately local context, e.g. as a coordinate subject, since the NP is a newly mentioned entity.

6. *Orkem and Coates said last Wednesday that the two were considering a merger, through Orkem's British subsidiary, Orkem Coatings U.K. Ltd.*

Thus, the lexical structures provide a rich set of schemas for argument mapping and semantic inferencing, as well as directed presuppositions for discontinuous semantic relations.

One final and important note about lexical structures and paradigmatic behavior. The seed information for these structures is largely derivable from machine-readable dictionaries. For example, a dictionary definition for *merger* (from the Longman Dictionary of Contemporary English is "the joining of 2 or more companies or firms" with subject code FINANCE. This makes the task of automatic construction of a robust lexicon for NLP applications a very realizable goal (cf. Boguraev (1991) and Wilks *et al.* (1991)).

4. Induction of Semantic Relations from Syntactic Forms

From discussion in the previous section, it should be clear that such paradigmatic information would be helpful if available. In this section, we present preliminary results indicating the feasability of learning LCPs from corpora, both tagged and untagged. Imagine being able to take the V-O pairs such as those given in section 2.1, and then applying semantic tags to the verbs which are appropriate to the role they play for that object (i.e. induction of the qualia roles for that noun). This is in fact the type of experiment reported on in Anick and Pustejovsky (1990). Here we apply a similar technique to a much larger corpus, in order to induce the *agentive* role for nouns. That is, the semantic predicate associated with bringing about the object.

In this example we look at the behavior of noun phrases and the prepositional phrases that follow them. In particular, we look at the co-occurrence of nominals with *between*, *with*, and *to*. Table 1 shows results of the conflating verb/noun plus preposition patterns. The percentage shown indicates the ratio of the particular collocation to the key word. Mutual information (MI) statistics for the two words in collocation are also shown. What these results indicate is that induction of semantic type from conflating syntactic patterns is possible. Based on the semantic types for these prepositions, the syntactic evidence suggests that there is a symmetric relation between the arguments in the following two patterns:

a. `Z with y` $= \lambda x \exists R_Z, y[R_Z(x,y) \wedge R_Z(y,x)]$

b. `Z between x and y` $=$ $\exists R_Z, x, y[R_Z(x,y) \wedge R_Z(y,x)]$

We then take these results and, for those nouns where the association ratios for `N with` and `N between` are similar, we pair them with the set of verbs governing these "NP PP" combinations in corpus, effectively partitioning the original V-O set into [+agentive] predicates and [-agentive] predicates. If our hypothesis is correct, we expect that verbs governing nominals collocated with a *with*-phrase will be mostly those predicates referring to the agentive quale of the nominal. This is because the *with*-phrase is unsaturated as a predicate, and acts to identify the agent of the verb as its argument. This is confirmed by our data, shown below.

Verb-Object Pairs with Prep = to

19	`form/V venture/O`
3	`announce/V venture/O`
3	`enter/V venture/O`
2	`discuss/V venture/O`
1	`be/V venture/O`
1	`abandon/V venture/O`
1	`begin/V venture/O`
1	`complete/V venture/O`
1	`negotiate/V venture/O`
1	`start/V venture/O`
1	`expect/V venture/O`

Conversely, verbs governing nominals collocating with a *between*-phrase will not refer to the agentive since the phrase is saturated already. Indeed, the only verb occurring in this position with any frequency is the copula *be*, namely with the following counts: `12 be/V venture/O`. Thus, week semantic types can be induced on the basis of syntactic behavior. In Pustejovsky et al (1991), we discuss how this general technique compares to somewhat different but related approaches described in Smadja (1991) and Zernik and Jacobs (1991).

5. Conclusion

We contend that using lexical semantic methods to guide lexical knowledge acquisition from corpora can yield structured thesaurus-like information in a form amenable for use within information retrieval applications. The work reported here illustrates the applicability of this approach for several important classes of nominals. Future work includes refining the discovery procedures to reduce misses and false alarms and extending the coverage of the lexical semantics component to allow the testing of such techniques on a greater range of terms. Finally, we are applying the results of the analysis within the context of data extraction for IR, to test their effectiveness as indexing and retrieval aids. Much of what we have outlined is still programmatic, but we believe that the approach to extracting information from corpora making use of lexical semantic information is a fruitful one and an area definitely worth exploring.

Acknowledgement

This research was supported by DARPA contract MDA904-91-C-9328. I would like to thank Scott Waterman for his assistance in preparing the statistics used here. I would also like to thank Scott Waterman, Federica Busa, Peter Anick, and Sabine Bergler for useful discussion.

References

1. Anick, P. and J. Pustejovsky (1990) "An Application of Lexical Semantics to Knowledge Acquisition from Corpora," *Proceedings of the 13th International Conference of Computational Linguistics*, August 20-25, 1990, Helsinki, Finland.

2. Bergler, S. (forthcoming) *The Evidential Analysis of Reported Speech*, Ph.D. Computer Science Department,Brandeis University.

3. Boguraev, B. "Building a Lexicon: The Contribution of Computers", in B. Boguraev (ed.), Special issue on computational lexicons, *International Journal of Lexicography*, **4(3)**, 1991.

4. Boguraev, B. and J. Pustejovsky (1990) "Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design," *Proceedings of the 13th International Conference of Computational Linguistics*, August 20-25, 1990, Helsinki, Finland.

5. Church, K. and Hanks, P., (1990) "Word Association Norms, Mutual Information and Lexicogra-

phy". *Computational Linguistics* Vol. 16(1).

6. Church, K. and D. Hindle (1990)" Collocational Constraints and Corpus-Based Linguistics." In *Working Notes of the AAAI Symposium: Text-Based Intelligent Systems*, 1990.
7. Grishman, R., L. Hirschman, N. Nhan (1986) "Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments." *Computational Linguistics*, Vol. 12, Number 3, pp. 205-215.
8. Hindle, D. and M. Rooth, "Structural Ambiguity and Lexical Relations", *Proceedings of the ACL*, 1991.
9. Levin, B. (1989) "The Lexical Organization of the English Verb", ms. to appear University of Chicago Press.
10. Pustejovsky, J. (1991) "The Generative Lexicon," *Computational Linguistics*, 17.4, 1991.
11. Pustejovsky, J. (forthcoming) *The Generative Lexicon: A Theory of Computational Lexical Semantics*, MIT Press, Cambridge, MA.
12. Pustejovsky, J. and P. Anick (1988) "The Semantic Interpretation of Nominals",*Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary.
13. Pustejovsky, J., S. Bergler, and P. Anick (1991) "Lexical Semantic Techniques for Corpus Analysis," (submitted to) *Computational Linguistics.*
14. Pustejovsky, J. and M. Rooth (in preparation) "Type Coercive Environments in Corpora".
15. Smadja, F. (1991) "Macro-coding the lexicon with co-occurrence knowledge," in Zernik (ed) *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, LEA, Hillsdale, NJ, 1991.
16. Wilks, Y., D. C. Fass, C.-M. Guo, J. E. McDonald, T. Plate, and B. M. Slator (1991) "Providing Machine Tractable Dictionary Tools," *Machine Translation* 5, 1991.
17. Zernik, U. (1989) "Lexicon Acquisition: Learning from Corpus by Exploiting Lexical Categories." *Proceedings of IJCAI 89.*
18. Zernik, U. and P. Jacobs (1990) "Tagging for learning: Collecting thematic relations from corpus." *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland.

	Word	Word	Word
Word	+ *to*	+ *with*	+ *between*
	(%)/MI	(%)/MI	(%)/MI
agreement	.117	.159	.028
	1.512	3.423	3.954
announcement	.010	.003	0
	-.918	-.409	*n/a*
barrier	.215	0	.030
	2.117	*n/a*	4.046
competition	.019	.028	.021
	-.269	1.701	3.666
confrontation	.029	.283	.074
	.141	4.000	4.932
contest	.052	.052	.039
	.715	2.323	4.301
contract	.066	.060	.002
	.947	2.463	1.701
deal	.028	.193	.004
	.086	3.616	2.015
dialogue	0	.326	.152
	n/a	4.140	5.644
difference	.017	.009	.348
	-.410	.638	6.474
expansion	.013	.007	0
	-.666	.381	*n/a*
impasse	0	.064	.096
	n/a	2.520	5.192
interactions	0	0	.250
	n/a	*n/a*	6.141
market	.013	.006	.000
	-.637	.240	-.500
range	.005	.002	.020
	-1.533	-.618	3.663
relations	.009	.217	.103
	-1.017	3.736	5.254
settlement	.013	.091	.012
	-.626	2.868	3.142
talks	.029	.218	.030
	.138	3.740	4.043
venture	.032	.105	.035
	.226	3.008	4.185
war	.010	.041	.015
	-.937	2.079	3.372

Table 1: Mutual information for noun/verb + preposition patterns.