Acquiring Predicate-Argument Mapping Information from Multilingual Texts

Chinatsu Aone, Douglas McKee

Systems Research and Applications (SRA)
2000 15th Street North
Arlington, VA 22201
aonec@sra.com, mckeed@sra.com

Abstract

This paper discusses automatic acquisition of predicate-argument mapping information from multilingual texts. The lexicon of our NLP system abstracts the language-dependent portion of predicate-argument mapping information from the core meaning of verb senses (i.e. *semantic concepts* as defined in the knowledge base). We represent this mapping information in terms of cross-linguistically generalized mapping types called *situation types* and word sense-specific *idiosyncrasies*. This representation has enabled us to automatically acquire predicate-argument mapping information, specifically situation types and idiosyncrasies, for verbs in English, Spanish, and Japanese texts.

1 Introduction

Lexicons for a natural language processing (NLP) system that perform syntactic and semantic analysis require more than purely syntactic (e.g. part-of-speech information) and semantic information (e.g. a concept hierarchy). Language understanding requires mapping from syntactic structures into conceptual representation (henceforth predicate-argument mapping), while language generation requires the inverse mapping. That is, grammatical functions in the syntactic structures (e.g. subject, object, etc.) should be mapped to thematic roles in the semantic structures (e.g. agent, theme, etc.).

In this paper, we discuss how we acquire such predicate-argument mapping information from multilingual texts automatically (cf. Zernik and Jacobs work on collecting thematic roles [20]). As discussed in Aone and Mckee [1], the lexicon of our NLP system abstracts the language-dependent portion of predicate-argument mapping information from the core meaning of verb senses (i.e. *semantic concepts* as defined in the knowledge base). We represent this mapping information in terms of cross-linguistically generalized mapping types called *situation types* and word sense-specific *idiosyncrasies*. This representation has enabled us to automatically acquire predicate-argument mapping information, specifically situation types and idiosyncrasies, for verbs in English, Spanish, and Japanese texts.

In the following sections, we first describe how we represent the predicate-mapping information. Then, we discuss how we acquire situation type and idiosyncrasy information automatically from multilingual texts and show some results.

2 Predicate-Argument Mapping Representation

Each lexical sense of a verb in our lexicon encodes its default predicate-argument mapping type (i.e. situation type), any word-specific mapping exceptions (i.e. idiosyncrasies), and

	# of required NP or S arguments	default thematic roles	prohibited thematic roles
CAUSED-PROCESS	2	Agent Theme	-
PROCESS-OR-STATE	1	Theme	Agent
AGENTIVE-ACTION	1	Agent	-
INVERSE-STATE	2	Goal Theme	Agent

Table 1: Definitions of Situation Types

	English	Spanish	Japanese
CAUSED-PROCESS	kill	matar, mirar	korosu, miru
PROCESS-OR-STATE	die	morir	shibousuru
AGENTIVE-ACTION	look	bailar	odoru
INVERSE-STATE	see	ver	mieru

Table 2: Situation Types and Verbs in Three Languages

its semantic meaning (i.e. semantic concept) in addition to its morphological and syntactic information. In the following, we discuss these three levels in detail.

2.1 Situation Types

Each of a verb's lexical senses is classified into one of the four default predicate-argument mapping types called *situation types*. As shown in Table 1, situation types of verbs are defined by two kinds of information: 1) the number of subcategorized NP or S arguments and 2) the types of thematic roles which these arguments should or should not map to. Since this kind of information is applicable to verbs of any language, situation types are language-independent predicate-argument mapping types. Thus, in any language, a verb of type CAUSED-PROCESS has two arguments which map to AGENT and THEME in the default case (e.g. "kill"). A verb of type PROCESS-OR-STATE has one argument whose thematic role is THEME, and it does not allow AGENT as one of its thematic roles (e.g. "die"). An AGENTIVE-ACTION verb also has one argument but the argument maps to AGENT (e.g. "look"). Finally, an INVERSE-STATE verb has two arguments which map to THEME and GOAL; it does not allow AGENT for its thematic role (e.g. "see"). Examples from three languages are shown in Table 2.

Although verbs in different languages are classified into the same four situation types using the same definition, mapping rules which map grammatical functions (i.e. subject, object, etc.) in the syntactic structures[1] to thematic roles in the semantic structures may differ from one language to another. This is because languages do not necessarily express the same thematic roles with the same grammatical functions. This mapping information is *language-specific* (cf. Nirenburg and Levin [16]).

The default mapping rules for the four situation types are shown in Table 3. They are nearly identical for the three languages (English, Spanish, and Japanese) we have analyzed so far. The only difference is that in Japanese the THEME of an INVERSE-STATE verb is expressed by marking the object NP with a particle "-ga", which is usually a subject

[1] We use structures similar to LFG's f-structures.

		English/Spanish Mapping	Japanese Mapping
CAUSED-PROCESS	AGENT	(SURFACE SUBJECT)	(SURFACE SUBJECT)
	THEME	(SURFACE OBJECT)	(SURFACE OBJECT)
PROCESS-OR-STATE	THEME	(SURFACE SUBJECT)	(SURFACE SUBJECT)
AGENTIVE-ACTION	AGENT	(SURFACE SUBJECT)	(SURFACE SUBJECT)
INVERSE-STATE	GOAL	(SURFACE SUBJECT)	(SURFACE SUBJECT)
	THEME	(SURFACE OBJECT)	(SURFACE OBJECT) (PARTICLE "GA")

Table 3: Default Mapping Rules for Three Languages

marker (cf. Kuno [12]).[2] [3] So we add such information to the INVERSE-STATE mapping rule for Japanese. Generalization expressed in situation types has saved us from defining semantic mapping rules for each verb sense in each language, and also made it possible to acquire them from large corpora automatically.

This classification system has been partially derived from Vendler and Dowty's aspectual classifications [19, 9] and Talmy's lexicalization patterns [18]. For example, all AGENTIVE-ACTION verbs are so-called *activity* verbs, and so-called *stative* verbs fall under either INVERSE-STATE (if transitive) or PROCESS-OR-STATE (if intransitive). However, the situation types are *not* for specifying the semantics of aspect, which is actually a property of the whole sentence rather than a verb itself (cf. Krifka [11], Dorr [8], Moens and Steedman [15]). For instance, as shown below, the same verb can be classified into two different aspectual classes (i.e. activity and accomplishment) depending on the types of object NP's or existence of certain PP's.

(1) a. Sue drank wine for/*in an hour.
    b. Sue drank a bottle of wine *for/in an hour.
(2) a. Harry climbed for/*in an hour.
    b. Harry climbed to the top *for/in an hour.

Situation types are intended to address the issue of cross-linguistic predicate-argument mapping generalization, rather than the semantics of aspect.

2.2 Idiosyncrasies

Idiosyncrasies slots in the lexicon specify word sense-specific idiosyncratic phenomena which cannot be captured by semantic concepts or situation types. In particular, subcategorized pre/postpositions of verbs are specified here. For example, the fact that "look" denotes its THEME argument by the preposition "at" is captured by specifying idiosyncrasies. Examples of lexical entries with idiosyncrasies in English, Spanish and Japanese are shown in Figure 1. As discussed in the next section, we derive this kind of word-specific information automatically from corpora.

[2] There is a debate over whether the NP with "ga" is a subject or object. However, our approach can accommodate either analysis.

[3] The GOAL of some INVERSE-STATE verbs in Japanese can be expressed by a "ni" postpositional phrase. However, as Kuno [12] points out, since this is an idiosyncratic phenomenon, such information does not go to the default mapping rule.

```
(LOOK (CATEGORY . V)
      (SENSE-NAME . LOOK-1)
      (SEMANTIC-CONCEPT #LOOK#)
      (IDIOSYNCRASIES (THEME (MAPPING (LITERAL "AT"))))
      (SITUATION-TYPE AGENTIVE-ACTION))

(INFECTAR (CATEGORY . V)
          (SENSE-NAME . INFECTAR-1)
          (SEMANTIC-CONCEPT #INFECT#)
          (IDIOSYNCRASIES (THEME (MAPPING (LITERAL "CON" "DE")))
                          (GOAL (MAPPING (SURFACE OBJECT))))
          (SITUATION-TYPE CAUSED-PROCESS))

(NARU (CATEGORY . V)
      (SENSE-NAME . NARU-1)
      (SEMANTIC-CONCEPT #BECOME#)
      (IDIOSYNCRASIES (GOAL (MAPPING (LITERAL "TO" "NI"))))
      (SITUATION-TYPE PROCESS-OR-STATE))
```

Figure 1: Lexical entries for "look", "infectar", and "naru"

2.3 Semantic Concepts

Each lexical meaning of a verb is represented by a semantic concept (or frame) in our language-independent knowledge base, which is similar to the one described in Onyshkevych and Nirenburg [17]. Each verb frame has thematic role slots, which have two facets, TYPE and MAPPING. A TYPE facet value of a given slot provides a constraint on the type of objects which can be the value of the slot. In the MAPPING facets, we have encoded some cross-linguistically general predicate-argument mapping information. For example, we have defined that all the subclasses of #COMMUNICATION-EVENT# (e.g. #REPORT#, #CONFIRM#, etc.) map their sentential complements (SENT-COMP) to THEME, as shown below.

```
(#COMMUNICATION-EVENT#
      (AKO #DYNAMIC-SITUATION#)
      (AGENT (TYPE #PERSON# #ORGANIZATION#))
      (THEME (TYPE #SITUATION# #ENTITY#)
             (MAPPING (SENT-COMP T)))
      (GOAL  (TYPE #PERSON# #ORGANIZATION#)
             (MAPPING (P-ARG GOAL))))
```

2.4 Merging Predicate-Argument Mapping Information

For each verb, the information stored in the three levels discussed above is merged to form a complete set of mapping rules. During this merging process, the idiosyncrasies take precedence over the situation types and the semantic concepts, and the situation types over the semantic concepts. For example, the two derived mapping rules for "break"(i.e. one for "break" as in "John broke the window" and the other for "break" as in "The window broke") are shown in Figure 2. Notice that the semantic TYPE restriction and INSTRUMENT role stored in the knowledge base are also inherited at this time.

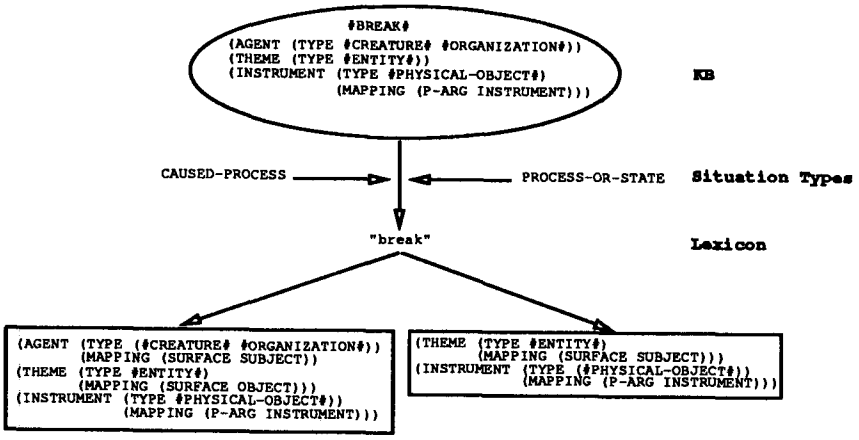

Figure 2: Information from the KB, the situation type, and the lexicon all combine to form two predicate-argument mappings for the verb "break."

3 Automatic Acquisition from Corpora

In order to expand our lexicon to the size needed for broad coverage and to be able to tune the system to specific domains quickly, we have implemented algorithms to automatically build multilingual lexicons from corpora. In this section, we discuss how the situation types and lexical idiosyncrasies are determined for verbs.

Our overall approach is to use simple robust parsing techniques that depend on a few language-dependent syntactic heuristics (e.g. in English and Spanish, a verb's object usually directly follows the verb), and a dictionary for part of speech information. We have used these techniques to acquire information from English, Spanish, and Japanese corpora varying in length from about 25000 words to 2.7 million words.

3.1 Acquiring Situation Type Information

We use two surface features to restrict the possible situation types of a verb: the verb's *transitivity rating* and its *subject animacy*.

The transitivity rating of a verb is defined to be the number of transitive occurrences in the corpus divided by the total occurrences of the verb. In English, a verb appears in the transitive when either:

- The verb is directly followed by a noun, determiner, personal pronoun, adjective, or wh-pronoun (e.g. "John owns a cow.")
- The verb is directly followed by a "THAT" as a subordinate conjunction (e.g. "John said that he liked llamas.")
- The verb is directly followed by an infinitive (e.g. "John promised to walk the dog.")
- The verb past participle is preceded by "BE," as would occur in a passive construction (e.g. "The apple was eaten by the pig.")

verb	occs	TR	SA	Pred. ST	Correct ST	Prepositional Idio
SUFFICE	8	0.6250	0.0000	(IS)	(IS)	
TIME	15	0.8333	1.0000	(CP IS)	(CP)	
TRAIN	20	1.0000	1.0000	(CP IS)	(CP PS)	at
WRAP	22	0.7222	0.6667	(CP IS)	(CP)	up over in with
SORT	25	0.4211	1.0000	(CP IS AA PS)	(CP AA)	out
UNITE	27	0.5833	1.0000	(CP IS AA PS)	(CP AA)	
TRANSPORT	28	0.8571	0.6667	(CP IS)	(CP)	
SUSTAIN	32	0.9062	0.6842	(CP IS)	(CP)	
SUBSTITUTE	33	0.7500	0.5000	(IS)	(CP PS)	for
TARGET	36	0.7778	0.8000	(CP IS)	(CP)	
STORE	36	0.9091	1.0000	(CP IS)	(CP)	on
STEAL	36	0.9167	0.6667	(CP IS)	(CP)	from
SHUT	36	0.2400	0.5000	(IS PS)	(CP PS)	up for
STRETCH	53	0.5278	0.5000	(IS PS)	(CP PS)	over into out from
STRIP	57	0.7609	0.8571	(CP IS)	(CP)	from into of
THREATEN	58	0.8793	0.4419	(IS)	(CP IS)	over
WEAR	61	0.8033	0.6667	(CP IS)	(IS)	over
TREAT	77	0.8052	0.8000	(CP IS)	(CP)	as
TERMINATE	79	0.9726	1.0000	(CP IS)	(CP PS)	
WEIGH	81	0.2069	0.5294	(IS PS)	(CP PS)	on with into
TEACH	82	0.7794	0.6875	(CP IS)	(CP)	at
SURROUND	85	0.8000	0.6667	(CP IS)	(CP)	
TOTAL	97	0.0515	0.2759	(PS)	(CP PS)	at
VARY	112	0.1354	0.0294	(IS PS)	(CP PS)	from over
WAIT	130	0.1923	1.0000	(CP IS AA PS)	(AA)	for up
SPEAK	139	0.1667	0.7500	(CP IS AA PS)	(AA CP)	out at up
SURVIVE	146	0.4754	0.3846	(IS PS)	(IS PS)	
UNDERSTAND	180	0.6946	0.8684	(CP IS)	(IS)	
SURGE	188	0.0182	0.3125	(PS)	(PS)	
SUPPLY	188	0.7176	0.8571	(CP IS)	(CP)	with
SIT	199	0.0625	0.7027	(AA PS)	(AA PS)	on with at out in up
TEND	200	0.8594	0.4340	(IS)	(CP IS)	
BREAK	219	0.4771	0.5000	(IS PS)	(CP PS)	up into out
WRITE	243	0.4637	0.9123	(CP IS AA PS)	(CP AA)	off
WATCH	268	0.7069	0.8462	(CP IS)	(CP)	out over
SUCCEED	277	0.5379	0.8899	(CP IS AA PS)	(CP PS)	
STAY	300	0.2156	0.6604	(CP IS AA PS)	(PS)	out up on with at
STAND	310	0.2841	0.7237	(CP IS AA PS)	(PS CP AA)	up at as out on
TELL	368	0.8054	0.8101	(CP IS)	(CP)	
SPEND	445	0.3823	0.8125	(CP IS AA PS)	(CP)	on over
SUPPORT	454	0.8486	0.5370	(IS)	(CP IS)	
SUGGEST	570	0.7782	0.5918	(IS)	(CP IS)	
TURN	852	0.3418	0.5891	(IS PS)	(CP PS)	out into up over
START	890	0.3474	0.6221	(CP IS AA PS)	(CP PS)	with off out
LOOK	1084	0.1718	0.6520	(CP IS AA PS)	(AA PS)	at into for up
THINK	1227	0.7602	0.9237	(CP IS)	(CP)	
TRY	1272	0.7904	0.8743	(CP IS)	(CP)	
WANT	1659	0.8559	0.8787	(CP IS)	(IS)	
USE	2211	0.8416	0.7725	(CP IS)	(CP)	
TAKE	2525	0.7447	0.5933	(IS)	(CP IS)	over off out into up

Table 4: Automatically Derived Situation Type and Idiosyncrasy Data

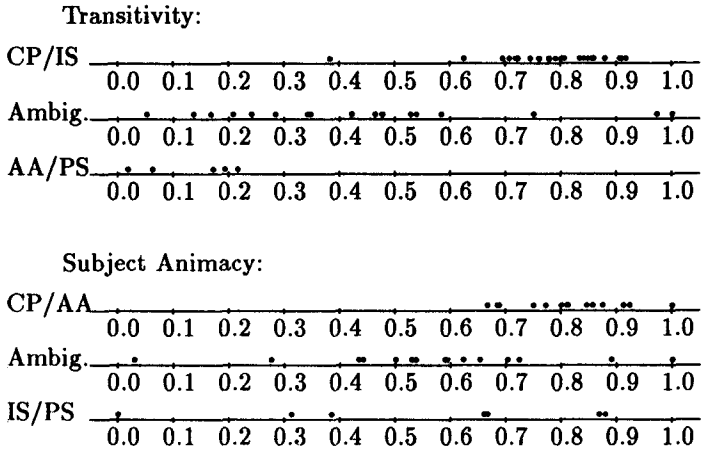

Figure 3: This graph shows the accuracy of the Transitivity and Subject Animacy metrics.

For Spanish, we use a very similar algorithm, and for Japanese, we look for noun phrases with an object marker "-wo" near and to the left of the verb. A high transitivity is correlated with CAUSED-PROCESS and INVERSE-STATE while a low transitivity correlates with AGENTIVE-ACTION and PROCESS-OR-STATE. Table 4 shows 50 verbs and their calculated transitivity rating. Figure 3 shows that for all but one of the verbs that are unambiguously transitive the transitivity rating is above 0.6. The verb "spend" has a transitivity rating of 0.38 because most of its direct objects are numeric dollar amounts. Phrases which begin with a number are not recognized as direct objects, since most numeric amounts following verbs are adjuncts as in "John ran 3 miles."

We define a verb's subject animacy to be the number of times the verb appears with an animate subject over the total occurrences of the verb where we identified the subject. Any noun or pronoun directly preceding a verb is considered to be its subject. This heuristic fails in cases where the subject NP is modified by a PP or relative clause as in "The man under the car wore a red shirt." We have only implemented this metric for English. The verb's subject is considered to be animate if it is any one of the following:

- A personal pronoun ("it" and "they" were excluded, since they may refer back to inanimate objects.)
- A proper name
- A word under "agent" or "people" in WordNet (cf. [14])
- A word that appears in a MUC-4 template slot that can be filled only with humans (cf. [7])

Verbs that have a low subject animacy cannot be either CAUSED-PROCESS or AGENTIVE-ACTION, since the syntactic subject must map to the AGENT thematic

role. A high subject animacy does not correlate with any particular situation type, since several stative verbs take only animate subjects (e.g. perception verbs).

The predicted situation types shown in Figure 3 were calculated with the following algorithm:

1. Assume that the verb can occur with every situation type.
2. If the transitivity rating is greater than 0.6, then discard the AGENTIVE-ACTION and PROCESS-OR-STATE possibilities.
3. If the transitivity rating is below 0.1, then discard the CAUSED-PROCESS and INVERSE-STATE possibilities.
4. If the subject animacy is below 0.6, then discard the CAUSED-PROCESS and AGENTIVE-ACTION possibilities.

We are planning several improvements to our situation type determination algorithms. First, because some stative verbs can take animate subjects (e.g. perception verbs like "see", "know", etc.), we sometimes cannot distinguish between INVERSE-STATE or PROCESS-OR-STATE and CAUSED-PROCESS or AGENTIVE-ACTION verbs. This problem, however, can be solved by using algorithms by Brent [3] or Dorr [8] for identifying stative verbs.

Second, verbs ambiguous between CAUSED-PROCESS and PROCESS-OR-STATE (e.g. "break", "vary") often get inconclusive results because they appear transitively about 50% of the time. When these verbs are transitive, the subjects are almost always animate and when they are intransitive, the subjects are nearly always inanimate. We plan to recognize these situations by calculating animacy separately for transitive and intransitive cases.

3.2 Acquiring Idiosyncratic Information

We automatically identify likely pre/postpositional argument structures for a given verb by looking for pre/postpositions in places where they are likely to attach to the verb (i.e. within a few words to the right for Spanish and English, and to the left for Japanese). When a particular pre/postposition appears here much more often than chance (based on either Mutual Information or a chi-squared test [5, 4]), we assume that it is a likely argument. A very similar strategy works well at identifying verbs that take sentential complements by looking for complementizers (e.g. "that", "to") in positions of likely attachment. Some English examples are shown in Tables 4 and 5, and Spanish examples are shown in Tables 6 and 7. The details of the exact algorithms used for English are contained in McKee and Maloney [13]. Areas for improvement include distinguishing between cases where a verb takes a prepositional arguments, a prepositional particle, or a common adjunct.

4 Conclusion

We have automatically built lexicons with predicate-argument mapping information from English, Spanish and Japanese corpora. These lexicons have been used for several multilingual data extraction applications (cf. Aone *et al.* [2]) and a prototype Japanese-English

word	possible clausal complements
know	THATCOMP
vow	THATCOMP, TOCOMP
eat	–
want	TOCOMP
resume	INGCOMP

Table 5: English Verbs which Take Complementizers

verb	MI with "que"
indicar	9.3
señalar	8.7
estimar	8.6
calcular	7.7
precisar	7.7
anunciar	7.7

Table 6: Spanish Verbs which Take Complementizers

verb	preposition	MI between verb and preposition
luchar	contra	12.4
unir	contra	8.9
vacunar	contra	8.9
cifrar	sobre	9.6
consultar	sobre	9.6
pasar	sobre	8.6
acordar	con	10.8
contar	con	10.3
relacionar	con	9.7
notificar	en	8.7
ocurrir	en	8.0
encontrar	en	7.8

Table 7: Spanish Verbs that Take Prepositional Arguments

machine translation system. The algorithms presented here have minimized our lexical acquisition effort considerably.

Currently we are investigating ways in which thematic role slots of verb frames and semantic type restrictions on these slots can be derived automatically from corpora (cf. Dagan and Itai [6], Hindle and Rooth [10], Zernik and Jacobs [20]) so that knowledge acquisition at all three levels of predicate-argument mapping can be automated.

References

[1] Chinatsu Aone and Doug McKee. Three-Level Knowledge Representation of Predicate-Argument Mapping for Multilingual Lexicons. In *AAAI Spring Symposium Working Notes on Building Lexicons for Machine Translation*, 1993.

[2] Chinatsu Aone, Doug McKee, Sandy Shinn, and Hatte Blejer. SRA: Description of the SOLOMON System as Used for MUC-4. In *Proceedings of Fourth Message Understanding Conference (MUC-4)*, 1992.

[3] Michael Brent. Automatic Semantic Classification of Verbs from Their Syntactic Contexts: An Implemented Classifier for Stativity. In *Proceedings of the 5th European ACL Conference*, 1991.

[4] Kenneth Church and William Gale. Concordances for Parallel Text. In *Proceedings of the Seventh Annual Conference of the University of Waterloo Centre for the New OED and Text Research: Using Corpora*, 1991.

[5] Kenneth Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), 1990.

[6] Ido Dagan and Alon Itai. Automatic Acquisition of Constraints for the Resolution of Anaphora References and Syntactic Ambiguities. In *Proceedings of the 13th International Conference on Computational Linguistics*, 1990.

[7] Defense Advanced Research Projects Agency. *Proceedings of Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann Publishers, 1992.

[8] Bonnie Dorr. A Parameterized Approach to Integrating Aspect with Lexical-Semantics for Machine Translation. In *Proceedings of 30th Annual Meeting of the ACL*, 1992.

[9] David Dowty. *Word Meaning and Montague Grammar*. D. Reidel, 1979.

[10] Donald Hindle and Mats Rooth. Structural Ambiguity and Lexical Relations. In *Proceedings of 29th Annual Meeting of the ACL*, 1991.

[11] Manfred Krifka. Nominal Reference, Temporal Construction, and Quantification in Event Semantics. In R. Bartsch et al., editors, *Semantics and Contextual Expressions*. Foris, Dordrecht, 1989.

[12] Susumu Kuno. *The Structure of the Japanese Language*. MIT Press, 1973.

[13] Doug McKee and John Maloney. Using Statistics Gained from Corpora in a Knowledge-Based NLP System. In *Proceedings of The AAAI Workshop on Statistically-Based NLP Techniques*, 1992.

[14] George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[15] Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2), 1988.

[16] Sergei Nirenburg and Lori Levin. Syntax-Driven and Ontology-Driven Lexical Semantics. In *Proceedings of ACL Lexical Semantics and Knowledge Representation Workshop*, 1991.

[17] Boyan Onyshkevych and Sergei Nirenburg. Lexicon, Ontology and Text Meaning. In *Proceedings of ACL Lexical Semantics and Knowledge Representation Workshop*, 1991.

[18] Leonard Talmy. Lexicalization Patterns: Semantic Structure in Lexical Forms. In Timothy Shopen, editor, *Language Typology and Syntactic Descriptions*. Cambridge University Press, 1985.

[19] Zeno Vendler. *Linguistics in Philosophy*. Cornell University Press, 1967.

[20] Uri Zernik and Paul Jacobs. Tagging for Learning: Collecting Thematic Relations from Corpus. In *Proceedings of the 13th International Conference on Computational Linguistics*, 1990.