The Semantics of Collocational Patterns for Reporting Verbs

Sabine Bergler
Computer Science Department
Brandeis University
Waltham, MA 02254
e-mail: sabine@chaos.cs.brandeis.edu

Abstract

One of the hardest problems for knowledge extraction from machine readable textual sources is distinguishing entities and events that are part of the main story from those that are part of the narrative structure. Importantly, however, reported speech in newspaper articles explicitly links these two levels. In this paper, we illustrate what the lexical semantics of reporting verbs must incorporate in order to contribute to the reconstruction of story and context. The lexical structures proposed are derived from the analysis of semantic collocations over large text corpora.

I Motivation

We can distinguish two levels in newspaper articles: the pure information, here called *primary information*, and the meta-information, which embeds the primary information within a perspective, a belief context, or a modality, which we call *circumstantial information*. The distinction is not limited to, but is best illustrated by, reported speech sentences. Here the matrix clause or *reporting clause* corresponds to the circumstantial information, while the complement (whether realized as a full clause or as a noun phrase) corresponds to primary information. For tasks such as knowledge extraction it is the primary information that is of interest. For example in the text of Figure 1 the matrix clauses (italicized) give the circumstantial information of the *who*, *when* and *how* of the reporting event, while *what* is reported (the primary information) is given in the complements.

The particular reporting verb also adds important information about the manner of the original utterance, the preciseness of the quote, the temporal relationship between matrix clause and complement, and more. In addition, the source of the original information provides information about the reliability or credibility of the primary information. Because the individual reporting verbs differ slightly but importantly in this respect, it is the lexical semantics that must account for such knowledge.

US Advising Third Parties on Hostages

(R1) *The Bush administration continued to insist yesterday that* (C1) it is not involved in negotiations over the Western hostages in Lebanon, (R2) *but acknowledged that* (C2) US officials have provided advice to and have been kept informed by "people at all levels" who are holding such talks.

(C3) "There's a lot happening, and I don't want to be discouraging," (R3) *Marlin Fitzwater, the president's spokesman, told reporters.* (R4) *But Fitzwater stressed that* (C4) he was not trying to fuel speculation about any impending release, (R5) *and said* (C5) there was "no reason to believe" the situation had changed.

(A1) Nevertheless, it appears that it has. ...

Figure 1: Boston Globe, March 6, 1990

We describe here a characterization of influences which the reporting clause has on the interpretation of the reported clause without fully analyzing the reported clause. This approach necessarily leaves many questions open, because the two clauses are so intimately linked that no one can be analyzed fully in isolation. Our goal is, however, to show a minimal requirement on the lexical semantics of the words involved, thereby enabling us to attempt a solution to the larger problems in text analysis.

The lexical semantic framework we assume in this paper is that of the *Generative Lexicon* introduced by Pustejovsky [Pustejovsky89]. This framework allows

us to represent explicitly even those semantic collocations which have traditionally been assumed to be presuppositions and not part of the lexicon itself.

II Semantic Collocations

Reporting verbs carry a varying amount of information regarding time, manner, factivity, reliability etc. of the original utterance. The most unmarked reporting verb is *say*. The only presupposition for *say* is that there was an original utterance, the assumption being that this utterance is represented as closely as possible. In this sense *say* is even less marked than *report*, which in addition specifies an adressee (usually implicit from the context.)

The other members in the semantic field are set apart through their semantic collocations. Let us consider in depth the case of *insist*. One usage can be found in the first part of the first sentence in Figure 1, repeated here as (1).

1 *The Bush administration continued to insist yesterday that it is not involved in negotiations over the Western hostages in Lebanon.*

The lexical definition of *insist* in the Longman Dictionary of Contemporary English (LDOCE) [Procter78] is

insist 1 to declare firmly (when opposed)

and in the Merriam Webster Pocket Dictionary (MWDP) [Woolf74]:

insist to take a resolute stand: PERSIST.

The opposition, mentioned explicitly in LDOCE but only hinted at in MWDP, is an important part of the meaning of *insist*. In a careful analysis of a 250,000 word text base of TIME magazine articles from 1963 (TIMEcorpus) [Bergler90a] we confirmed that in every sentence containing *insist* some kind of opposition could be recovered and was supported by some other means (such as emphasis through word order etc.). The most common form of expressing the opposition was through negation, as in (1) above.

In an automatic analysis of the 7 million word corpus containing Wall Street Journal documents (WSJC) [Bergler90b], we found the distribution of patterns of opposition reported in Figure 2. This analysis shows that of 586 occurrences of *insist* throughout the WSJC, 109 were instances of the idiom *insisted on* which does not subcategorize for a clausal complement. Ignoring these occurrences for now, of the remaining 477 occurrences, 428 cooccur

Keywords	Occ	Comments
insist	586	occurrences throughout the corpus
insist on	109	these have been cleaned by hand and are actually occurrences of the idiom *insist on* rather than accidental co-occurrences.
insist & but	117	occurrences of both *insist* and *but* in the same sentence
insist & negation	186	includes *not* and *n't*
insist & subjunctive	159	includes *would*, *could*, *should*, and *be*
insist & but & neg.	14	
insist & but & on	12	
insist & but & subj.	8	

Figure 2: Negative markers with *insist* in WSJC

with such explicit markers of opposition as *but* (selecting for two clauses that stand in an opposition), *not* and *n't*, and subjunctive markers (indicating an opposition to factivity). While this is a rough analysis and contains some "noise", it supports the findings of our careful study on the TIMEcorpus, namely the following:

2 A propositional opposition is implicit in the lexical semantics of *insist*.

This is where our proposal goes beyond traditional collocational information, as for example recently argued for by Smadja and McKeown [Smadja&McKeown90]. They argue for a flexible lexicon design that can accomodate both single word entries and collocational patterns of different strength and rigidity. But the collocations considered in their proposal are all based on word cooccurrences, not taking advantage of the even richer layer of semantic collocations made use of in this proposal. Semantic collocations are harder to extract than cooccurrence patterns—the state of the art does not enable us to find semantic collocations automatically[1]. This paper however argues that if we take advantage of lexical paradigmatic behavior underlying the lexicon, we can at least achieve semi-automatic extraction of semantic collocations (see also Calzolari and Bindi (1990)

[1]But note the important work by Hindle [Hindle90] on extracting *semantically* similar nouns based on their substitutability in certain verb contexts. We see his work as very similar in spirit.

and Pustejovsky and Anick (1990) for a description of tools for such a semi-automatic acquisition of semantic information from a large corpus).

Using *qualia structure* as a means for structuring different semantic fields for a word [Pustejovsky89], we can summarize the discussion of the lexical semantics of *insist* with a preliminary definition, making explicit the underlying opposition to the assumed context (here denoted by ψ) and the fact that *insist* is a reporting verb.

3 (Preliminary Lexical Definition)

insist(A,B)
[Form: Reporting Verb*]*
[Telic: utter(A,B) & $\exists\psi$: opposed(B,ψ)*]*
[Agentive: human(A)*]*

III Logical Metonymy

In the previous section we argued that certain semantic collocations are part of the lexical semantics of a word. In this section we will show that reporting verbs as a class allow *logical metonymy* [Pustejovsky91] [Pustejovsky&Anick88]. An example can be found in (1), where the metonymy is found in the subject NP. *The Bush administration* is a compositional object of type *administration*, which is defined somewhat like (4).

4 (Lexical Definition)

administration
[Form: + plural
part of: institution*]*
[Telic: execute(x, orders(y)),
where y is a high official
in the specific institution]
[Constitutive: + human
executives,
officials,...*]*
[Agentive: appoint(y, x)*]*

In its formal role at least an *administration* does not fulfill the requirements for making an utterance— only in its constitutive role is there the attribute *[+ human]*, allowing for the metonymic use.

Although metonymy is a general device — in that it can appear in almost any context and make use of associations never considered before[2] — a closer look at the data reveals, however, that metonymy as used in newspaper articles is much more restricted and systematic, corresponding very closely to *logical metonymy* [Pustejovsky89].

Not all reporting verbs use the same kind of metonymy, however. Different reporting verbs select for different semantic features in their source NPs. More precisely, they seem to distinguish between a single person, a group of persons, and an institution. We confirmed this preference on the TIMEcorpus, extracting automatically all the sentences containing one of seven reporting verbs and analyzing these data by hand. While the number of occurrences of each reporting verb was much too small to deduce the verb's lexical semantics, they nevertheless exhibited interesting tendencies.

Figure 3 shows the distribution of the degree of animacy. The numbers indicate percent of total occurrence of the verb, i.e. in 100 sentences that contain *insist* as a reporting verb, 57 have a single person as their source.

	person	group	instit.	other
admit	64%	19%	14%	2%
announce	51%	10%	31%	8%
claim	35%	21%	38%	6%
denied	55%	17%	17%	11%
insist	57%	24%	16%	3%
said	83%	6%	4%	8%
told	69%	7%	8%	16%

Figure 3: Degree of Animacy in Reporting Verbs

The significance of the results in Figure 3 is that semantically related words have very similar distributions and that this distribution differs from the distribution of less related words. *Admit, denied* and *insist* then fall in one category that we can call here informally [-inst], *said* and *told* fall in [+person], and *claim* and *announce* fall into a not yet clearly marked category [other]. We are currently implementing statistical methods to perform similar analyses on WSJC. We hope that the impreciseness of an automated analysis using statistical methods will be counterbalanced by very clear results.

The TIMEcorpus also exhibited a preference for one particular metonymy, which is of special interest for reporting verbs, namely where the name of a country, of a country's citizens, of a capital, or even of the building in which the government resides stands for the government itself. Examples are *Great Britain/ The British/ London/ Buckingham Palace announced....* Figure 4 shows the preference of the re-

[2] As the well-known example *The ham sandwich ordered another coke.* illustrates.

porting verbs for this metonymy in subject position. Again the numbers are too small to say anything about each lexical entry, but the difference in preference is strong enough to suggest it is not only due to the specific style of the magazine, but that some metonymies form strong collocations that should be reflected in the lexicon. Such results in addition provide interesting data for preference driven semantic analysis such as Wilks' [Wilks75].

Verb	percent of all occurrences
admit	5%
announce	18%
claim	25%
denied	33%
insist	9%
said	3%
told	0%

Figure 4: *Country, countrymen,* or *capital* standing for the *government* in subject position of 7 reporting verbs.

IV A Source NP Grammar

The analysis of the subject NPs of all occurrences of the 7 verbs listed in Figure 3 displayed great regularity in the TIMEcorpus. Not only was the logical metonymy discussed in the previous section pervasive, but moreover a fairly rigid semantic grammar for the source NPs emerged. Two rules of this semantic grammar are listed in Figure 5.

```
source ⟶
        [quant] [mod] descriptor ["," name ","] |
        [descriptor | ((a | the) mod)] [mod] name |
        [inst 's | name 's] descriptor [name] |
        name "," [a | the] [relation prep] descriptor |
        name "," [a | the] name 's (descriptor
        | relation) |
        name "," free relative clause
descriptor ⟶
        role |
        [inst] position |
        [position (for | of)] [quant] inst
```

Figure 5: Two rules in a semantic grammar for source NPs

The grammar exemplified in Figure 5 is partial — it only captures the regularities found in the TIMEcorpus. Source NPs, like all NPs, can be adorned with modifiers, temporal adjuncts, appositions, and relative clauses of any shape. The important observation is that these cases are very rare in the corpus data and must be dealt with by general (i.e. syntactic) principles.

The value of a specialized semantic grammar for source NPs is that it provides a powerful interface between lexical semantics, syntax, and compositional semantics. Our source NP grammar compiles different kinds of knowledge. It spells out explicitly that logical metonymy is to be expected in the context of reporting verbs. Moreover, it *restricts* possible metonymies: the *ham sandwich* is not a typical source with reporting verbs. The source grammar also gives a likely *ordering* of pertinent information as roughly COUNTRY|LOCATION ALLEGIANCE INSTITUTION POSITION NAME.
This information defines essentially the *schema* for the representation of the source in the knowledge extraction domain.

We are currently applying this grammar to the data in WSJC in order to see whether it is specific to the TIMEcorpus. Preliminary results were encouraging: The adjustments needed so far consisted only of small enhancements such as adding locative PPs at the end of a descriptor.

V LCPs—Lexical Conceptual Paradigms

The data that lead to our source NP grammar was essentially collocational material: We extracted the subject NPs for a set of verbs, analyzed the lexicalization of the source and generalized the findings[3]. In this section we will justify why we think that the results can properly be generalized and what impact this has on the representation in the lexicon.

It has been noted that dictionary definitions form a — usually shallow — hierarchy [Amsler80]. Unfortunately explicitness is often traded in for conciseness in dictionaries, and conceptual hierarchies cannot be automatically extracted from dictionaries alone. Yet for a computational lexicon, explicit dependencies in the form of lexical inheritance are crucial [Briscoe&al.90] [Pustejovsky&Boguraev91]. Following Anick and Pustejovsky (1990), we argue that lexical items having related, paradigmatic syntactic behavior enter into the same *lexical conceptual paradigm*. This states that items within an LCP will have a set of syntactic realization patterns for how the

[3] A detailed report on the analysis can be found in [Bergler90a]

word and its conceptual space (e.g. presuppositions) are realized in a text. For example, reporting verbs form such a paradigm. In fact the definition of an individual word often stresses the difference between it and the closest synonym rather than giving a constructive (decompositional) definition (see LDOCE).[4] Given these assumptions, we will revise our definition of *insist* in (3). We introduce an LCP (i.e. semantic type), REPORTING VERB, which spells out the core semantics of reporting verbs. It also makes explicit reference to the source NP grammar discussed in Section IV as the default grammar for the subject NP (in active voice). This general template allows us to define the individual lexical entry concisely in a form close to normal dictionary definitions: deviations and enhancements as well as restrictions of the general pattern are expressed for the individual entry, making a comparison between two entries focus on the differences in entailments.

5 (Definition of Semantic Type)

REPORTING VERB
[Form: ∃A,B,C,D: utter(A,B)
& hear(C,B)
& utter(C, utter(A,B))
& hear(D,utter(C, utter(A,B)))*]*
[Constitutive: SUBJECT: type:SourceNP,
COMPLEMENT *]*
[Agentive: AGENT*(C)*, COAGENT*(A)]*

6 (Lexical Definition)

insist(A,B)
[Form: REPORTING VERB*]*
[Telic: ∃ψ: opposed(B,ψ)*]*
[Constitutive: MANNER: vehement*]*
[Agentive: [-inst]*]*

A related word, *deny*, might be defined as 7.

7 (Lexical Definition)

deny(A,B)
[Form: REPORTING VERB*]*
[Telic: ∃ψ: negate(B,ψ)*]*
[Agentive: [-inst]*]*

(6) and (7) differ in the quality of their opposition to the assumed proposition in the context, ψ: *insist* only specifies an opposition, whereas *deny* actually negates that proposition. The entries also reflect their common preference not to participate in the metonymy that allows *institutions* to appear in subject position. Note that *opposed* and *negate* are not assumed to be primitives but decompositions; these predicates are themselves decomposed further in the lexicon.

Insist (and other reporting verbs) "inherit" much structural information from their semantic type, i.e. the LCP REPORTING VERB. It is the semantic type that actually provides the constructive definition, whereas the individual entries only define refinements on the type. This follows standard inheritance mechanisms for inheritance hierarchies [Pustejovsky&Boguraev91] [Evans&Gazdar90].

Among other things the LCP REPORTING VERB specifies our specialized semantic grammar for one of its constituents, namely the subject NP in non-passive usage. This not only enhances the tools available to a parser in providing semantic constraints useful for constituent delimiting, but also provides an elegant way to explicitly state which logical metonymies are common with a given class of words[5].

VI Summary

Reported speech is an important phenomenon that cannot be ignored when analyzing newspaper articles. We argue that the lexical semantics of reporting verbs plays an important part in extracting information from large on-line text bases.

Based on extensive studies of two corpora, the 250,000 word TIMEcorpus and the 7 million word Wall Street Journal Corpus we identified that *semantic collocations* must be represented in the lexicon, expanding thus on current trends to include syntactic collocations in a word based lexicon [Smadja&McKeown90].

We further discovered that *logical metonymy* is pervasive in subject position of reporting verbs, but that reporting verbs differ with respect to their preference for different kinds of logical metonymy. A careful analysis of seven reporting verbs in the TIMEcorpus suggested that there are three features that divide the reporting verbs into classes according to the preference for metonymy in subject position, namely whether the subject NP refers to the source as a single person, a group of people, or an institution.

The analysis of the source NPs of seven reporting verbs further allowed us to formulate a specialized se-

[4] The notion of LCPs is of course related to the idea of semantic fields [Trier31].

[5] Grimshaw [Grimshaw79] argues that verbs also select for their complements on a semantic basis. For the sake of conciseness the whole issue of the form of the complement and its semantic connection has to be omitted here.

mantic grammar for source NPs, which constitutes an important interface between lexical semantics, syntax, and compositional semantics used by an application program. We are currently testing the completeness of this grammar on a different corpus and are planning to implement a noun phrase parser.

We have imbedded the findings in the framework of Pustejovsky's *Generative Lexicon* and *qualia theory* [Pustejovsky89] [Pustejovsky91]. This rich knowledge representation scheme allows us to represent explicitly the underlying structure of the lexicon, including the clustering of entries into semantic types (i.e. LCPs) with inheritance and the representation of information which was previously considered presuppositional and not part of the lexical entry itself. In this process we observed that the analysis of semantic collocations can serve as a measure of semantic closeness of words.

Acknowledgements: I would like to thank my advisor, James Pustejovsky, for inspiring discussions and many critical readings.

References

[Amsler80] Robert A. Amsler. *The Structure of the Merriam-Webster Pocket Dictionary.* PhD thesis, University of Texas, 1980.

[Anick&Pustejovsky90] Peter Anick and James Pustejovsky. Knowledge acquisition from corpora. In *Proceedings of the 13th International Conference on Computational Linguistics*, 1990.

[Briscoe&al.90] Ted Briscoe, Ann Copestake, and Branimir Boguraev. Enjoy the paper: Lexical semantics via lexicology. In *Proceedings of the 13th International Conference on Computational Linguistics*, 1990.

[Bergler90a] Sabine Bergler. Collocation patterns for verbs of reported speech—a corpus analysis on the time Magazine corpus. Technical report, Brandeis University Computer Science, 1990.

[Bergler90b] Sabine Bergler. Collocation patterns for verbs of reported speech—a corpus analysis on The Wall Street Journal. Technical report, Brandeis University Computer Science, 1990.

[Calzolari&Bindi90] Nicoletta Calzolari and Remo Bindi. Acquisition of lexical information from a large textual italian corpus. In *Proceedings of the 13th International Conference on Computational Linguistics*, 1990.

[Evans&Gazdar90] Roger Evans and Gerald Gazdar. The DATR papers. Cognitive Science Research Paper CSRP 139, School of Cognitive and Computing Sciences, University of Sussex, 1990.

[Grimshaw79] Jane Grimshaw. Complement selection and the lexicon. *Linguistic Inquiry*, 1979.

[Hindle90] Donald Hindle. Noun classification from predicate-argument structures. In *Proceedings of the Association for Computational Linguistics*, 1990.

[Pustejovsky&Anick88] James Pustejovsky and Peter Anick. The semantic interpretation of nominals. In *Proceedings of the 12th International Conference on Computational Linguistics*, 1988.

[Pustejovsky&Boguraev91] James Pustejovsky and Branimir Boguraev. A richer characterization of dictionary entries. In B. Atkins and A. Zampolli, editors, *Computer Assisted Dictionary Compiling: Theory and Practice.* Oxford University Press, to appear.

[Pustejovsky89] James Pustejovsky. Issues in computational lexical semantics. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 1989.

[Pustejovsky91] James Pustejovsky. Towards a generative lexicon. *Computational Linguistics*, 17, 1991.

[Procter78] Paul Procter, editor. *Longman Dictionary of Contemporary English.* Longman, Harlow, U.K., 1978.

[Smadja&McKeown90] Frank A. Smadja and Kathleen R. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the Association for Computational Linguistics*, 1990.

[Trier31] Jost Trier. *Der deutsche Wortschatz im Sinnbezirk des Verstandes: Die Geschichte eines sprachlichen Feldes. Bandl.* Heidelberg, 1931.

[Wilks75] Yorick Wilks. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6, 1975.

[Woolf74] Henry B. Woolf, editor. *The Merriam-Webster Dictionary.* Pocket Books, New York, 1974.