Class Specificity and the Lexical Encoding of Participant Information

Jean-Pierre Koenig
Center for Cognitive Science and Linguistics Department, University at Buffalo, The State University of New York

and

Gail Mauner and Breton Bienvenue
Center for Cognitive Science and Psychology Department, University at Buffalo, The State University of New York

Published online November 13, 2001

It is commonly assumed across the language sciences that some semantic participant information is lexically encoded and some is not. In this article, we propose that semantic obligatoriness and verb class specificity are criteria which influence whether semantic information is lexically encoded. We present a comprehensive survey of the English verbal lexicon and two continuation studies which confirm that both factors play a role in the lexical encoding of participant information.

Key Words: verb; arguments; adjuncts; sentence processing; lexical semantics; instruments; locations.

It is commonly assumed across the language sciences that only some semantic participant information is lexically encoded. The distinction is critical to grammatical theories since the syntax of "basic" sentences is driven by their main verbs' arguments (see Bresnan, 1982; Chomsky, 1981; Foley & Van Valin, 1984; Pollard & Sag, 1987; among others). It is also important to research on human sentence processing. It has been claimed that the interpretation of filler-gap dependencies is sensitive to arguments but not adjuncts (e.g., Boland & Tanenhaus, 1990; Boland, Tanenhaus, Garnsey, & Carlson, 1995). Similarly, attachment ambiguity resolution has been argued to be sensitive to the argument status of phrases which can be ambiguously attached (e.g., Schütze & Gibson, 1999; Speer & Clifton, 1998). Finally, the distinction between arguments and adjuncts might place a limit on the lexical interpretation of implicit semantic dependents (Mauner & Koenig, 2000; Mauner, Tanenhaus, & Carlson, 1995) in that only lexically encoded implicit participant information is expected to play a role in the immediate representation that readers form for sentences.

This research was supported (in part) by Research Grant 1 R01 MH60133-01 from the National Institute of Mental Health, National Institutes of Health; and by grants from the Center for Cognitive Science at the University at Buffalo. We gratefully acknowledge Tom Graves, Cori Grimm, Leana Longley, and Alissa Melinger for their work on the English verbal lexicon project; and Leana Longley and Cameron Stelmach for assistance in data collection and analysis.

Address correspondence and reprint requests to Jean-Pierre Koenig, Department of Linguistics, 609 Baldy Hall, University at Buffalo, Buffalo, NY 14260-1030. E-mail: jpkoenig@acsu.buffalo.edu.
Unfortunately, despite the large number of extant proposals, no set of necessary and sufficient criteria has yet been proposed as the basis for the distinction between arguments and adjuncts. In this article, we propose two criteria that jointly determine the argument status of participant information. We provide evidence for this hypothesis from a comprehensive survey of the English verbal lexicon and two sentence continuation studies.

Because the literature ascribes different meanings to the same terms, we begin with a few terminological pointers. We call semantic participants the set of entities (in a broad sense of the term) which are included in a situation. Semantic participants include the situation’s agents, patients, and other traditional roles, but also locations and times (what Langacker, 1987, calls ‘‘settings’’). We call lexically encoded semantic participants arguments and nonlexically encoded semantic participants adjuncts. We call syntactically active arguments, those lexically encoded participants which can license a semantically dependent syntactic expression (complements, adverbs, unexpressed subjects of rationale clauses, and so forth). Thus, the unexpressed agent participant of a passive verb is syntactically active, whereas the same unexpressed participant of a middle verb is syntactically inactive (see Williams, 1987; Mauner & Koenig, 2000). We reserve the terms complement and modifier for syntactically co-occurring constituents that are lexically encoded or not, respectively.

Existing criteria for distinguishing between arguments and adjuncts have come under fire of late (Schütze, 1995; and especially Miller, 1997). For reasons of space, we cannot go into the details of these scholars’ arguments and simply refer the reader to their work as well as Koenig, Mauner, and Bienvenue (2001). Because it is often assumed that argumenthood reduces to obligatory syntactic expression, we simply note here that overt syntactic expression is not a possible criterion for the lexical encoding of semantic participant information for two reasons. First, it does not cover optional syntactic constituents which can play a role in sentence processing. Linguists have argued that passive participles such as sold in (1a), but not middle verbs such as sell in (1b), include a syntactically unexpressed agent argument (see Williams, 1987), as suggested by the fact that only the former can serve as the antecedent for the unexpressed subject of a rationale clause. Mauner et al. (1995) and Mauner and Koenig (2000) provide corroborating experimental evidence for this conclusion. The undisputed argument status of the unexpressed agent of the passive participle sold demonstrates that argumenthood cannot be equated with syntactic obligatoriness or overt syntactic expression.

(1a) The vase was sold to collect money for the charity.

(1b)* The vase sold to collect money for the charity.

Second, arguments need not even be able to co-occur with a verb (pace Mauner, 1996), as shown by the behavior of Spanish reflexive passives exemplified in (2). These passives behave like English passives in licensing a rationale clause, but by contrast to what happens in the English case, the agent cannot be expressed. On the assumption that licensing of the unexpressed subject of a rationale clause is evidence for the lexical encoding of an agent argument, the Spanish facts suggest that the lexical encoding of arguments does not require overt expressibility of that information. This is important, since it suggests that argumenthood cannot reduce entirely to textual co-occurrence information, as suggested by MacDonald, Pearlmutter, and Seidenberg (1994). Additionally, as pointed out by Keenan (1985) among others, the Spanish data can probably be generalized, since, cross-linguistically, passive verbs or verb phrases often do not allow the expression of the verb’s agent.
(2) El florero se vendió rápidamente (*por la organización)
The vase sold rapidly (*by the organization)

para recaudar fondos para los niños.
to gather funds for the children.

to gather money for children.

Even if some linguistic criteria do reliably distinguish participant information that is lexically encoded (i.e., arguments) from that which is not (i.e. adjuncts)—e.g., the do so test (Lakoff & Ross, 1976) or constraints on extraction (Huang, 1982)—they presuppose speakers’ knowledge of the distinction between arguments and adjuncts (see Koenig et al., 2001). The absence of clear, observable properties on which to base the distinction between arguments and adjuncts raises the following question: What observable properties do language learners and users rely on to determine which participant information is included in a lexical item’s representation in sentences like (3)?

(3) Mary chopped wood with a hatchet for her sick neighbor on the porch last Sunday.

We define lexically encoded information as that information which is accessed immediately upon recognition of a word. Furthermore, we propose that the lexical encoding of participant information is based on two criteria. Lexically encoded participant information is obligatory; that is, in more technical terms, it is entailed to hold of the class of situations denoted by a word, and it is relatively specific to the relevant verb. Both properties are criteria which language learners can directly observe (by observing, for example, the systematic co-occurrence of participants in described situations or the fact that this participant information is not shared by the overwhelming majority of verbs). It thus provides a basis for learning the distinction between arguments and adjuncts.

The relevance of obligatoriness is illustrated in the contrast between chop in (3) and eat in (4).

(4) Mary ate her fish with a spoon.

Whereas chop requires that all situations it felicitously describes include an instrument, eat does not; it merely allows the inclusion of an instrument. The use of semantic obligatoriness as a criterion for lexical encoding of participant information is not new; it has been proposed as a necessary condition for argumenthood as early as Dowty (1982). Although it has been discarded by scholars such as Bresnan (1982), we believe her data only argue against the use of obligatoriness as constituting a necessary and sufficient condition on argumenthood. In our model, obligatoriness is supplemented by verb specificity.

The contrast between the PPs with a hatchet and on the porch in (3) illustrates verb specificity. For the moment we will assert that the obligatory presence of an instrument is true of a restricted set of verbs (informally, those verbs which describe actions performed by an agent who needs to employ an instrument), while the obligatory presence of a location in which the event occurs—what we call event location—is true of most verbs. The instrument participant information is thus more specific to chop than is the event location information. Similarly, and in a more subtle manner, the presence of a location which locates the entire event described by a verb [e.g., on the porch in sentence (3)] is required by many more verbs than the presence of a location which merely locates one participant in the event described by the verb.
[e.g., in the drawer in sentence (5)]. We call this latter kind of locations participant location.

(5) Marc put his socks in the drawer.

The role we attribute to verb specificity in determining argumenthood stems from two considerations. First, at a fine-grained level, the more specific a piece of semantic information is, the greater its contribution to the individuation of a word’s meaning, that is, to what truly characterizes the conceptual content of that word. The forceful contact and incision of the instrument into the patient, for example, separates chop from other verbs which require an instrument, such as draw. If we assume that the conceptual content of a word is retrieved upon its recognition, we expect, all else being equal, such specific information to be associated more closely with a word than nonspecific information that does nothing to individuate one verb from another, such as the fact that an event took place somewhere. Second, at a more coarse-grained level, recent research suggests that many syntactic processes target semantically defined verb classes and that individual verbs are associated with such classes (see Davis & Koenig, 2000; Goldberg, 1995; Levin, 1993; Pinker, 1989). Thus, some verbs which require an instrument display the valence alternation illustrated in (6).

(6a) Bill hit the wall with a stick.
(6b) Bill hit the stick against the wall.

The presence of such semantically determined alternations shows that participant information at the coarser grained level of semantic roles such as instrument (or agent, patient, and so forth) is also relevant to linguistic and, more particularly, syntactic processes. We therefore expect that upon recognition of a verb, those semantically determined verb classes of which the verb is a member, as well as the semantic information which partly defines those classes, are activated.

The coarse- and fine-grained effects of verb specificity can be seen in our model of the organization of lexical knowledge. We assume that lexical knowledge can be described as a multidimensional hierarchy of categories, along the lines of a traditional semantic network (see Brachman & Schmolze, 1985; Collins, Quillian, & Ross, 1970; Quillian, 1968). Each category can include a combination of syntactic, semantic, or morphological information. Words which are members of these categories are linked to the most specific categories of which they are members and each category can itself be linked to more general categories and so forth. This view of lexical knowledge underlies most directly syntactic research within Head-driven Phrase Structure Grammar (see Pollard & Sag, 1987, 1994, for an introduction, and Koenig, 1999, for a detailed analysis of lexical knowledge along these lines). But, it is not crucial to our analysis that the criteria which jointly determine lexical encoding of participant information be implemented in the mental lexicon in the form of inheritance networks. As Rumelhart and Todd (1993) show, an organization of knowledge very similar, if not isomorphic, to semantic networks may emerge out of a distributed representational schema. What is crucial for our hypothesis regarding lexical encoding of participant information is that upon reading or hearing a word, addressees access a vast amount of semantic and syntactic information which is not encapsulated, but rather shared across words. In such a model, lexical encoding means that the information accessed or activated upon the recognition of a word includes information about categories to which the word is linked. Lexical encoding of participant information thus reduces to the semantic categories accessed or activated upon word recognition. The situation is illustrated informally for Causes and Affected entities, the most undisputed classes of arguments, in Fig. 1.

Each node in this figure represents a cluster of semantic, syntactic, or graphemic/
phonological information. The semantic representation of the transitive use of *break* is represented by the node labeled *break*\textsubscript{caus}. It inherits schematic participant information from two general semantic classes, the class of situations which include a cause, represented by the node labeled *cause* and the class of situations which include an affected entity, represented by the node labeled *affected*. The use of dashed lines indicates the presence of intermediary nodes between the two categories. The word *break* also inherits syntactic properties from the node labeled *transitive-verb*, which summarizes information which characterizes the syntactic category of transitive verbs. Because this syntactic category is itself associated with the semantic nodes representing situations which include causes and affected entities, as argued by various scholars (see Goldberg, 1995, for a review), the nodes *cause* and *affected* are also activated, since the category *transitive-verb* is activated.

Together, our hypothesis regarding the semantic bases of argumenthood and our model of the organization of lexical knowledge predicts that obligatory participants corresponding to arguments are associated with restricted classes of verbs while obligatory participants corresponding to adjuncts are associated with most verbs. We refer to this as the Semantic Selectivity Hypothesis (hereafter SSH). We tested the SSH in two ways: A quantitative survey of the English verbal lexicon and two sentence continuation studies.

**QUANTITATIVE SURVEY OF THE ENGLISH VERBAL LEXICON**

At a coarse-grained level of semantic granularity, the SSH predicts a sharp drop in the percentage of verbs which require a given participant role between lexically encoded and nonlexically encoded participant information. To test this hypothesis, we selected agent properties as a benchmark for arguments. Agents are the most frequent and indisputable examples of arguments and can therefore be used to determine a cutoff point for class selectivity.

To estimate the proportion of verbs which required agent properties, we had pairs of raters assess, for the verbs in the MRC psycholinguistic database (Coltheart, 1981), whether they described a situation in which particular participant properties were required (3909 verbs of the approximately 5500 listed verbs were known to all raters). Three agent properties, corresponding to Dowty (1991) and Davis and Koenig (2000), were assessed as follows:

1. The class of adjuncts also include semantically optional participants, such as beneficiary. For reasons of space, we do not discuss those in this article.
ENCODING OF PARTICIPANT INFORMATION

TABLE 1
Percentages of Verbs Judged to Require
Agent Properties

<table>
<thead>
<tr>
<th>Verb type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causal Force</td>
<td>29.8</td>
</tr>
<tr>
<td>Volition</td>
<td>23.6</td>
</tr>
<tr>
<td>Notion</td>
<td>14.2</td>
</tr>
</tbody>
</table>

1. Was one of the participants in the situations described by the verb volitionally involved (e.g., Martha in Martha jogged last night)?
2. Was one of the participants in the situations described by the verb causally affecting another participant (e.g., Marc in Marc finally cooked the fish)?
3. Had one of the participants in the situations described by the verb a mental representation of another participant (e.g., Marc in Marc thought of the beach)?

After engaging in a resolution process, percentages of interrater agreements were computed for verbs which required these participant properties. All interrater agreements were above 99%. We conducted separate ratings for the listed agent properties, since, as argued by Dowty (1991), the concept of agent is a cluster category.

The results of the survey are summarized in Table 1. The most important finding of this first survey is that even the most frequent agent property, cause, was restricted to less than 30% of the verbal lexicon. This suggests that arguments are true of a restricted class of verbs. Having established a benchmark for class selectivity, we chose as a test case the comparison between the categories of instruments, participant locations, event locations, and times. Event locations and times are widely agreed to constitute adjuncts and the SSH predicts that, by contrast to agent properties, the obligatory presence of these participants in described situations will not be specific to a restricted class of verbs. We also compared event locations and times to instruments because the argument status of instruments is controversial. Schütze (1995) and Van Valin and Lapolla (1997) argue that instruments pass most argument tests, whereas Carlson and Tanenhaus (1988) and Dowty (1989) argue that they are adjunctive. It is possible that the inconclusiveness of previous studies stems from treating instruments as a unified category and from not recognizing the existence of two classes of verbs, those which semantically allow, but do not require the presence of instruments in denoted situations [e.g., drink in (7b)] and those which do require their presence [e.g., carve in (7a)]. Only the latter can be arguments according to our hypothesis. Finally, we compared event locations and times to the seldom noticed category of participant locations (i.e., of locations which indicate where a participant in the event is or ends up) which are exemplified by the PPs in her notebook and in his desk in sentences (8a) and (8b) respectively. Participant locations behave like the category of obligatory instruments and are therefore predicted to pattern like obligatory instruments with respect to verb class restrictiveness.

(7a) The seventh-grader carved his name on the desk with a pocket knife.
(7b) The boy drank his soda with a straw.
(8a) Johanna wrote the address in her notebook, (while) in her office.
(8b) Bill hid many compromising pictures in his desk, (while) at school.

The procedure followed to assess the percentage of verbs which require this new set of participant properties was identical to that described for agent properties. For example, after the distinction between participant and event location was illustrated,
TABLE 2
Percentages of Verbs Judged to Require Semantic Roles

<table>
<thead>
<tr>
<th>Verb type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instruments</td>
<td>12</td>
</tr>
<tr>
<td>External locations</td>
<td>98.2</td>
</tr>
<tr>
<td>Participant locations</td>
<td>7</td>
</tr>
<tr>
<td>Time</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Raters were asked to assess for each verb in the database whether it semantically required or allowed the presence of a participant and event location in all situations it felicitously described. More precisely, raters judged whether there was one sense of the verb whose denotation required or allowed the presence of the relevant participant. A verb such as put, for example, was classified by our raters as requiring the presence of a participant location because any event it felicitously describes entails the presence of a location in which the theme ends up. By contrast, a verb such as push was classified as allowing, but not requiring, the presence of a participant location because, presumably, one can push on an entity without it moving and therefore ending up in a new location. In this second survey, 4142 verbs were known to all raters. Table 2 summarizes the results of this survey. All interrater agreements were above 93%.

As can be seen, there are many fewer verbs which require participant locations and instruments than there are which require event locations and times. Thus, obligatory instruments and participant locations pattern with agent properties in that they are true of restricted classes of verbs. This suggests that obligatory instruments and participant locations are arguments. Together, the results of our two surveys confirm the predictions of the SSH: Participant categories widely held to be adjunctives (event locations and time) are less restrictive (i.e., are required of a much larger percentage of verbs) than participant categories which are widely held to be arguments (e.g., various agent properties). Furthermore, our two surveys suggest that other participant properties (instruments and participant locations) are also arguments. Finally, they also highlight the fact that participant type cuts across semantic obligatoriness and that, therefore, participant categories cannot be used as indices of argumenthood. This observation goes a long way toward reconciling the contradictory results regarding the argument status of instruments.

THE PSYCHOLOGICAL REALITY OF THE SSH: TWO SENTENCE CONTINUATION STUDIES

To test the psychological validity of the SSH, we conducted two sentence continuation studies. The assumptions underlying these studies are that participant information which is lexically encoded is retrieved upon recognition of a word and that participant information that is activated is more likely to be used in providing a sentence continuation. Given these assumptions, we predicted that sentences with obligatory instrument verbs would elicit more instrument continuations than sentences with control verbs that merely allowed instrument participants; we also predicted that sentences with obligatory instrument or participant location verbs would elicit more instrument and participant location continuations respectively than other continuation types. (Since control verbs in the participant location study did not allow the inclusion of participant locations, a comparison between the percentages of participant location
completions following experimental and control verbs was not possible.) In the first study, participants provided continuations for 18 experimental sentences whose main verbs semantically require an instrument, e.g., (9a), and 36 control sentences whose main verbs do not semantically require, but allow, an instrument, e.g., (9b). In the second study, participants provided continuations for 31 experimental sentences whose main verbs semantically require a participant location, e.g., (10a), and 36 control sentences whose main verbs do not semantically require a participant location, e.g., (10b).

(9a) The farmer split the logs ____________________________
(9b) Dana ate a bowl of soup ____________________________

(10a) The gang member drew an obscene picture ________________________
(10b) The IRS audited Robert’s tax returns __________________________

Through a pilot study in which we coded for 14 categories, we determined that 7 and 8 of them were needed to classify continuations in studies 1 and 2, respectively. The remaining categories did not elicit significant numbers of continuations and were aggregated into a single Other category. The results of the two studies are represented in Figs. 2 and 3, respectively.

Three sets of $X^2$ analyses revealed that (1) verbs semantically requiring an instrument (e.g., split) or participant location (e.g., draw) elicited significantly more instrument and participant location completions respectively than expected by chance; (2) verbs semantically requiring an instrument or participant location elicited significantly more instrument and participant location completions respectively than any other type of completion; and (3) verbs that semantically require instruments (e.g, split) elicited significantly more instrument completions than verbs that allow, but do not require instruments (e.g., eat).

Together, these results corroborate our survey results: Arguments are true of re-

---

2 The single category of locations for the instrument study was split into two participant and event location subcategories in the location study.
stricted sets of verbs. Participant information which holds of a limited set of verbs was more accessible for expression than participant information which holds of almost all verbs.

Although these results support the SSH, a competing explanation is that the obtained differences in percentages of continuations are not due to a semantic contrast between experimental and control verbs or between purported arguments and adjuncts, but rather to the sheer surface frequency of PPs expressing the various participant properties. PPs which express locations or instruments merely co-occur more frequently with obligatory instruments and participant location verbs than with non-obligatory instruments and nonparticipant location verbs. To assess this hypothesis, we searched the Brown corpus for occurrences of our participant location and control verbs and tabulated the percentage of times these verbs occurred with various PPs.

Although most participant location verbs rarely occurred in the corpus, when they did, they very rarely co-occurred with a constituent expressing a participant location. More importantly, the correlation between frequency of co-occurrence in the corpus and percentage of participant location completions was not significant and in the wrong direction \( r^2 = -0.27, p > .5 \). This suggests that surface co-occurrence did not drive the effect we found in at least the participant location continuation study. Note that differences in the frequency of expression of syntactically optional semantic arguments in our continuation studies and in (nonlaboratory generated) ordinary texts are not that surprising. The fact that a verb lexically encodes a syntactically optional semantic participant, but another does not is no guarantee that speakers or writers will express that semantic participant more often with the former. Multiple factors influence the expression of syntactically optional semantic arguments, including (i) the verb-specific restrictiveness of the selectional constraints on fillers of that participant role (see Resnick, 1997) or (ii) the discourse role that this filler will play (see Givón, 1984). These other lexical and discourse factors can override any difference in rate of expression which the difference in lexical encoding would lead us to expect. By contrast, our experimental participants were explicitly asked to provide continuations and thereby express syntactically optional phrases. We hypothesize that in such a situation the above-mentioned factors for not expressing participant information
are severely attenuated such that the importance of the lexical encoding of that information becomes more important.

Interestingly enough, a similar analysis for the instrument continuation study did show a correlation between co-occurrence of a verb and a with-PP in the Brown corpus, on the one hand, and percentages of instrument continuations, on the other. One possible explanation for the existence of a correlation between textual co-occurrence and continuation percentages in the case of instruments, but not participant locations, is that many control items in the instrument continuation study included optional instrument verbs which describe situations which are very unlikely to involve an instrument (e.g., begin and overhear). By contrast, all situations described by control verbs in the location continuation study were required to include an event location. Given the limited corpus which we used to compute co-occurrence frequencies, this meant that instrument control verbs were quite unlikely to co-occur with with-phrases, and indeed, 16 of the 18 control verbs never co-occurred with a with-PP. By contrast, instrument verbs (which describe situations which must include an instrument) are much more likely to co-occur with with-PP at least some of the time, thus accounting for the correlation we found between continuation percentages and co-occurrence with with-PPs.

In this article, we have argued that lexical encoding of participant information reduces to two semantic criteria: (1) whether participant information is semantically obligatory and (2) whether participant information is specific to a verb or to a restricted verb class to which a verb belongs. We have shown through a comprehensive survey of the English verbal lexicon that participant roles that correspond to the traditional notion of argument as well as more controversial roles such as semantically obligatory instruments and participant locations do display class selectivity, as we predicted. The results of two sentence continuation studies support the psychological relevance of the distinction. Verbs which, according to our criteria, lexically encode an argument (either an instrument or participant location) are more likely to lead to continuations which express that argument than they are to continuations which express adjuncts. Also, continuations which express a given participant role are more likely to occur after verbs which lexically encode that participant role than after verbs which do not. While the current results support the psychological reality of the argument/adjunct distinction we propose, they do not show that readers use it on-line. We have conducted experiments comparing the processing of filler-gap dependencies of verbs which do and do not lexically encode an instrument which suggest that readers do use it on-line. Because of space limitations, we cannot repeat the details of these studies here (see Mauner, Bienvenue, & Koenig, 2000).

REFERENCES


