MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1160 June, 1989

LEXICAL CONCEPTUAL STRUCTURE AND GENERATION IN MACHINE TRANSLATION

Bonnie J. Dorr

Abstract: This report introduces an implemented scheme for generating target-language sentences using a compositional representation of meaning called *lexical conceptual structure*. Lexical conceptual structure facilitates two crucial operations associated with generation: lexical selection and syntactic realization. The compositional nature of the representation is particularly valuable for these two operations when semantically equivalent source- and target-language words and phrases are structurally or thematically divergent. For example, the English verb *to stab* may be translated as the composite Spanish form *dar cuchilladas a* (literally, *to knife* or *to give knife-wounds to*). To determine the correct lexical items and syntactic realization associated with the surface form in such cases, the underlying lexical-semantic forms are systematically mapped to the target-language syntactic structures. The model described constitutes a lexical-semantic extension to UNITRAN [Dorr, 1987], a syntactic-based translation system that is bidirectional between Spanish and English.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for this research has been provided by a grant from the Kapor Family Foundation and by NSF Grant DCR-85552543 under a Presidential Young Investigator's Award to Professor Robert C. Berwick. Useful guidance and commentary during this research were provided by Bob Berwick, Michael Brent, Bruce Dawson, Sandiway Fong, and Mike Kashket. This report is a revised version of a paper appearing in the Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, Ann Arbor, MI, 1989.

1 Introduction

This report describes an implemented generation system that matches the underlying conceptual structure of a sentence to the appropriate target-language lexical items and produces the structural realization of the target-sentence by means of syntactic mappings associated with these lexical items. This work represents a shift away from complex, language-specific, syntactic generation without entirely abandoning syntax. Furthermore, this work moves toward a model that employs a well-defined lexical conceptual representation without depending on situational context, expectations, or complex knowledge representations. Two crucial operations, *lexical selection* of target-language terms and *syntactic realization* of target-language forms, will be examined, and structural and thematic divergences that encumber these two operations will be discussed.

Consider the following example of translation from English to Spanish:

(1) I stabbed John ⇒ Yo di cuchilladas a Juan
(I gave knife-wounds to John)

Two properties of the system enable it to provide an appropriate translation for cases such as (1). The first is that the system relies on the notion of *compositionality* in order to select target-language terms. For example, because of the inherently compositional nature of the English source-language verb *stab*, the system is able to select the composite Spanish form *dar cuchilladas a* (literally, *to knife* or *to give knife-wounds to*) as the target-language equivalent. The second property of the system is that it relies on an abstraction of lexical-semantic information from syntactic information. For example, the system is able to choose the lexical item *dar* (literally, *give*) as the translation of *stab* without regard to its syntactic realization, and it is able to realize the phrase *a Juan* (literally, *to John*) in place of *John* without regard to its lexical-semantic structure.

Other generators for machine translation have been either syntactic-based (see [McDonald, 1987], [McKeown, 1985], and [Slocum, 1984]) or semantic-based (see [Cullingford, 1986], [Lytinen, 1987], [Nirenburg *et. al.*, 1987], and [Schank & Abelson, 1977]).[1] We will see that syntactic-based approaches are not adequate for translation in cases such as (1) since they do not take

[1]The reader should note that the division between syntactic and semantic approaches is not as clean-cut as implied here. For example, systems such as MUMBLE [McDonald, 1983, 1987] and TEXT [McKeown, 1985] are not entirely syntactic-based in that they use discourse and focus constraints to derive *messages* (*i.e.*, underlying representational forms); and systems such as SAM [Cullingford, 1981] and [Schank & Abelson, 1977] and MOPTRANS [Lytinen, 1987], which rely on the current situational context and expectations, are not entirely semantic-based since they take syntax into account for target-term positioning.

advantage of the lexical-semantic properties that aid the selection process; in addition, we will see that semantic-based approaches are not adequate for this example since they do not take advantage of syntactic information that aids the realization process.[2]

The next section describes two levels of description included in each lexical word entry: syntactic and lexical-semantic. Section 3 shows how these two levels aid the lexical selection and syntactic realization operations despite various structural and thematic divergences that arise during the generation process. Section 4 shows how these two operations are applied to example (1). Throughout the report, we will see that compositionality and syntactic/lexical-semantic abstraction are crucial to the model presented here.

2 Background for the Generation Scheme

The work of Jackendoff [1983] has been the primary influence on the design of UNITRAN's lexical-semantic generator. The representation adopted is *lexical conceptual structure* (henceforth LCS) as formulated by Hale and Laughren [1983] and Hale and Keyser [1986]. Each lexical entry has two levels of description: the first is the syntactic description (*i.e.*, θ-roles, category, and hierarchical and linear positioning of each argument associated with a lexical root word) and the second is a lexical-semantic description (*i.e.*, the LCS of the lexical root word).[3] For example, the lexical entry for the word *stab* is:

(2)
```
(DEF-ROOT-WORD (STAB)
  ;; Syntactic description
  :CAT (V)
  :INTERNAL ((Y THEME KNIFE-WOUND) (Z N GOAL))
         or ((Y THEME KNIFE-WOUND) (Z N GOAL)
             (U P INSTRUMENT SHARP-OBJECT INANIMATE))
  :EXTERNAL ((X N AGENT ANIMATE))
  ;; LCS description
  :LCS (CAUSE X (GO-POSS Y (TOWARD-POSS (AT-POSS Y Z)))
                (WITH-INSTR *HEAD* U)))
```

[2]The system described here is implemented in Common Lisp and is currently running on a Symbolics 3600 series machine. Because it translates one sentence at a time, it does not incorporate context or domain knowledge; thus, it cannot use discourse, situational expectations, or domain information in order to generate a sentence. Consequently, there are a number of capabilities found in systems such as MUMBLE, TEXT, SAM, and MOPTRANS, that cannot be reproduced here including external pronominal reference, paraphrasing, story telling, interactive question-answering, *etc.*

[3]It is possible to use a more general linking strategy that relates variables in the LCS with variables in the syntactic structure (*e.g.*, see [Jackendoff, 1989]. Such a strategy would allow structural positioning of arguments to be determined independent of the lexical entries. This possibility is investigated in [Dorr, 1989].

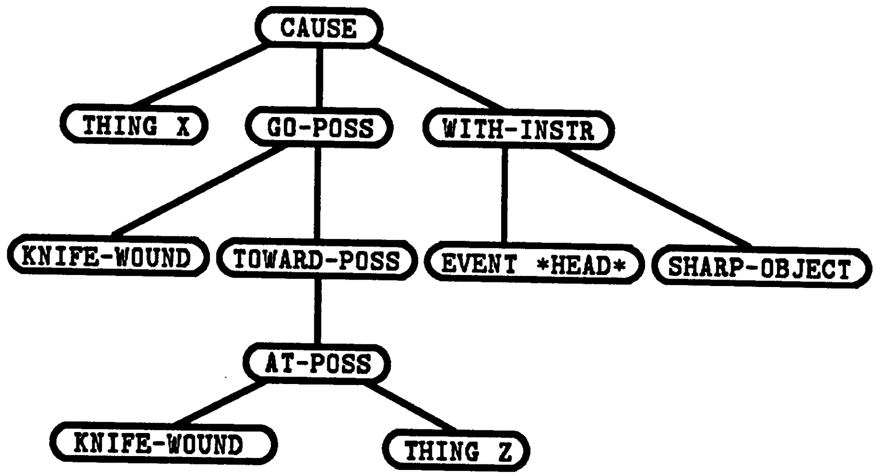

Figure 1: Underlying LCS for English *stab* and Spanish *dar* (*cuchilladas*)

The LCS description provides the meaning "THING X causes a possessional transfer of a knife wound THING Y to THING Z using a sharp object THING U as an instrument." Note that the instrument argument U is included in the LCS even though the source-language does not realize this argument in (1). Including this argument in the LCS allows flexibility in generating the target-language sentence, which may or may not require this argument to be realized. Thus, it would be possible to generate either *he stabbed the robber*, or *he stabbed the robber with a knife* (*scissors*, *poker*, *etc.*). The disjunction in the :INTERNAL slot of the root word definition allows for both possibilities. The *HEAD* symbol is a place-holder that points to the overall *stab* event (*i.e.*, the *stab* event is performed with a sharp object). Figure 1 shows the underlying LCS tree structure for the *stab* event.

The lexical-semantic primitives of the system will not be enumerated here. To summarize, I adopt Jackendoff's notions of EVENT and STATE; these are further specialized into such primitives as CAUSE, GO, BE, STAY, and LET. The specialized primitives are placed into Temporal, Locational, Possessional, Identificational, Circumstantial, Instrumental, Intentional, and Existential fields. For example, the primitive GO-POSS refers to a GO event in the Possessional field (*e.g.*, Beth received (= GO-POSS) the doll). If the GO event were placed in a Temporal field, it would become GO-TEMP (*e.g.*, the meeting went (= GO-TEMP) from 2:00 to 4:00). In addition to EVENTs and STATEs, there are also THINGs (*e.g.*, BOOK, PERSON, REFERENT *etc.*), PATHs (*e.g.*, TO, FROM, *etc.*), LOCATIONs and TIMEs (*e.g.*, HERE, TODAY, *etc.*), POSITIONs (*e.g.*, AT, WITH, *etc.*), PROPERTYs (*e.g.*, TIRED, HUNGRY, *etc.*), MANNERs (*e.g.*, FORCEFULLY, WELL, *etc.*), and INTENSIFIERs (*e.g.*, VERY, *etc.*). One difference between Jackendoff's representation and the one shown here is that the two-place predicate POSITION (*e.g.*, **AT**, **WITH**, *etc.*) is used instead of the one-place predicate **PLACE**; thus, the **KNIFE-WOUND** argument in figure 1 appears both internally and externally to the **AT-POSS** LCS node. Although the system uses only a small set of lexical-semantic primitives, the set is quite adequate for defining a potentially large set of words due to the compositional nature of LCS's. Furthermore, because the set is so small, the search space during the lexical-selection stage of generation is greatly reduced.

Given these two components of a lexical entry, a composed LCS can be constructed from the source-language parse tree (using the lexical-semantic description), and a target-language parse tree can then be generated from the composed LCS (using the syntactic description).[4] In the next section, we will see how this representation is used in the generation scheme.

3 Overview of the Generation Process

Two top-level generation procedures are activated after a source-language sentence has been parsed. The first is a lexical-semantic composition procedure that maps the source-language syntactic tree into an underlying composed LCS; the second is a syntactic generation routine that maps the underlying composed LCS into a target-language syntactic tree. The lexical-semantic composition task is implemented as a recursive procedure that converts a lexical word (henceforth referred to as the *head*) into its corresponding LCS, and then does the same for each of the arguments of that head. These LCS forms are then composed into a single LCS that underlies the source- and target-language sentences. The syntactic generation task is also a recursive procedure; it maps a node in the composed LCS to an appropriate target-language head, and then does the same for each of the arguments of that node. Each target-language head is then projected to its phrasal (or *maximal*) level and attached according to the positioning requirements of the lexical head that selects it.[5]

We return to our translation example shown in (1). Figure 2 shows how the LCS is composed from the parse tree for the *stab* event.[6] When the LCS-composition procedure is applied to the parse tree, the heads *I*,

[4]Although the examples in this report describe translation in one direction only, the composed LCS is actually a *pivot* (language-independent form) for translation in either direction.

[5]For discussion of projection to maximal level by the $\overline{X}$ component of the system, see [Dorr, 1987]. In a nutshell, **X-MAX** refers to the **XP** phrase that has a lexical head of category **X**.

[6]In this case, there is only one possible parse; however, if the structure were ambiguous, other possibilities would be displayed. The *e* elements under **C** and **I** are syntactic positions for which there is no overt lexical material.

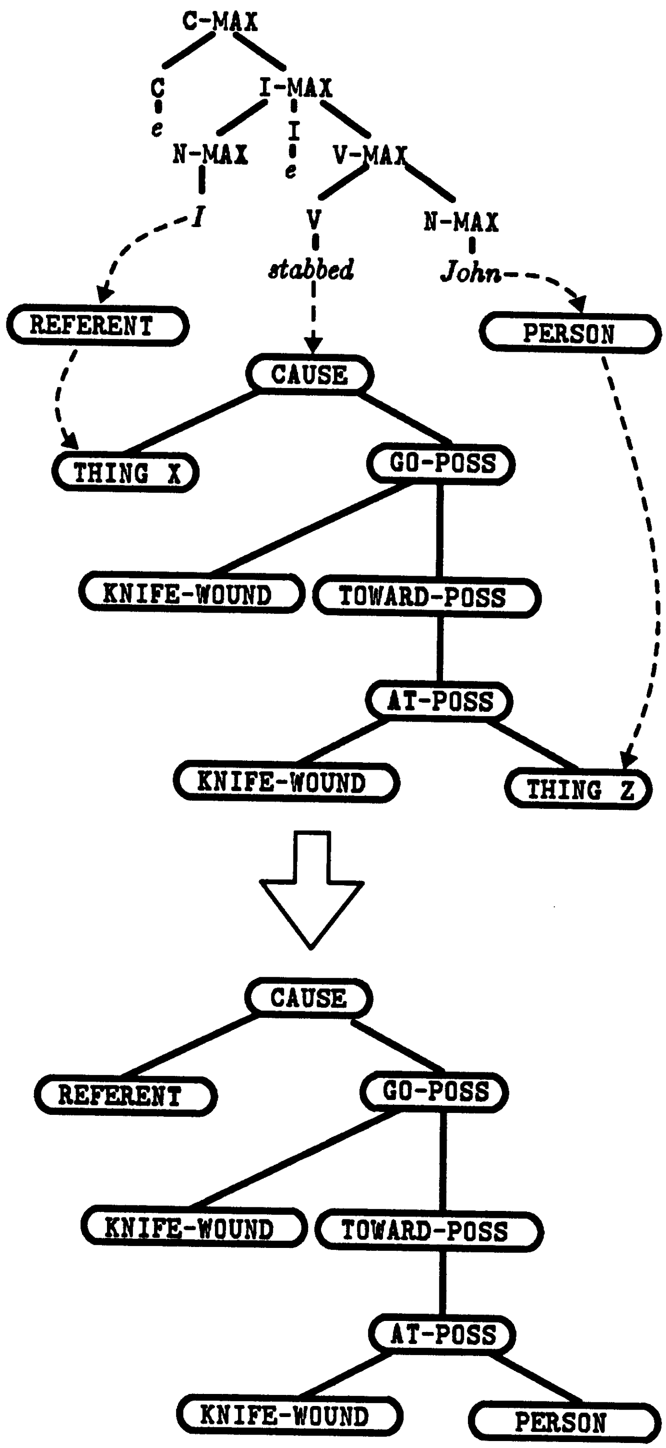

Figure 2: Composing the LCS for the *stab* Event

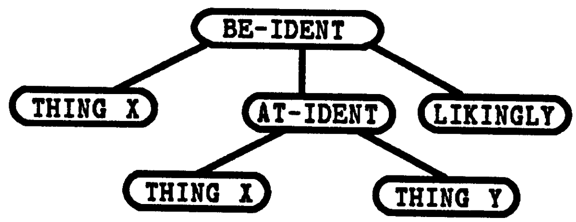

Figure 3: Underlying LCS for English *like* and Spanish *gustar*

stab, and *John* are isolated, and the corresponding LCS's are positioned according to the syntax-to-LCS mapping defined in the lexicon. Thus, the internal argument specification (Z N GOAL) in (2) maps the N-MAX projected from *John* to the variable Z. Similarly, the external argument specification (X N AGENT ANIMATE) maps the N-MAX projected from *I* to the variable X. The result is the composed LCS shown in figure 2.

Once the LCS has been composed, the syntactic generation component undertakes the tasks of *lexical selection* and *syntactic realization* to produce the target-language tree. We will now examine these two tasks in more detail before describing the process for the current example.

3.1 Lexical Selection: Thematic Divergence

Lexical selection is the task of choosing the target-language words that accurately reflect the meaning of the corresponding source-language words. One of the difficulties of this task is the fact that the equivalent source- and target-language forms are potentially thematically divergent. An example of thematic divergence shows up in the translation of the Spanish word *gustar* to the English word *like*. Although these two verbs are semantically equivalent, their argument structures are not identical: the subject of *like* (*I*) is the *theme* of the action, whereas the subject of *gustar* (*el libro*) is the *agent* of the action.[7] Thus, we have:

(3) Me gusta el libro (The book pleases me) ⇒ I like the book

In a syntactic-based scheme, the semantics of the verb *gustar* would be lost since the literal translation (*to please*) would be selected for the target-language verb. By contrast, a semantic-based system would generally be able

[7]In (3), the subject of the source-language sentence has freely inverted into post-verbal position, leaving behind a coindexed *pro* (empty pronominal element). Thus, the post-verbal subject is considered to be the external argument of the main verb. Free subject-inversion is a property of *pro-drop* languages (*i.e.*, languages such as Spanish, Italian, Hebrew, *etc.* that do not require a sentence to have a subject); this property is taken into account during syntactic parsing and generation. For further discussion of the principles and parameters underlying the parser, see [Dorr, 1987].

to make the correct lexical selection, but it might have difficulty with syntactic realization of the target-language arguments because it has no notion of syntactic argument divergence.

In the LCS approach, the underlying conceptual structure for *gustar* and *like* is identical (see figure 3), but the syntactic mappings associated with these two verbs are language-specific. The LCS underlying *gustar* and *like* reflects the fact that "`THING X` is in an identificational state `LIKINGLY` with respect to `THING Y`." However, the variables `X` and `Y` map to different syntactic positions for Spanish and English:

```
(4)   gustar:  :INTERNAL ((Y P THEME ANIMATE))
               :EXTERNAL ((X N AGENT))
      like:    :INTERNAL ((X N AGENT))
               :EXTERNAL ((Y N THEME ANIMATE))
```

Thus, the agent of the action becomes the subject (external argument) in Spanish, and the object (internal argument) in English.[8]

During syntactic generation, lexical selection of a target-language head involves matching the composed LCS to the appropriate lexical head in a target-language possibility set. For example, suppose the system is trying to select the appropriate target-language token for the composed LCS that corresponds to the source-language verb *gustar*. Several target heads (including *like*, *be*, and many others that use the `BE-IDENT` LCS) are selected as possible lexical possibilities. Each of these possibilities is then examined for a match: not only must the top-level LCS coincide, but all LCS's under the top-level LCS must also coincide. In general, there are two classes of LCS nodes that are taken into consideration during the matching process of lexical-selection. The more general nodes (*e.g.*, `BE-IDENT`, `AT-IDENT`, *etc.*) allow the matcher to determine the LCS class of the target-language term; the more specific nodes (*e.g.*, `LIKINGLY`, `FORCEFULLY`, *etc.*) are used for final convergence on a particular target-language term such as *like* as opposed to *love*, and *force* as opposed to *cause*.

In this example, the system determines that the *like* LCS is a match because it contains a `BE-IDENT` event whose arguments coincide with the arguments of the `BE-IDENT` in the composed LCS.[9] Figure 4 shows the translation process for this example.

---

[8]Notice also that the syntactic categories of the theme are not the same; this structural divergence shows up during syntactic realization, which will be discussed in section 3.2.

[9]There is still the question of what to do when the LCS-matching procedure does not adequately cut down the target-language possibilities. For example, there are many open-ended classes of words (in particular, noun-phrases, adjectives, and adverbs) that are not distinguishable by their LCS's. If the possibility list is still quite large (*i.e.*, more than two or three lexical items) after LCS-matching routines have finished the lexical selection process, a direct-mapping routine is used here instead for lexicalization. That is, certain lexical-items (*e.g.*, *me*, *I*, *John*, etc.) may be selected on the basis of a direct mapping

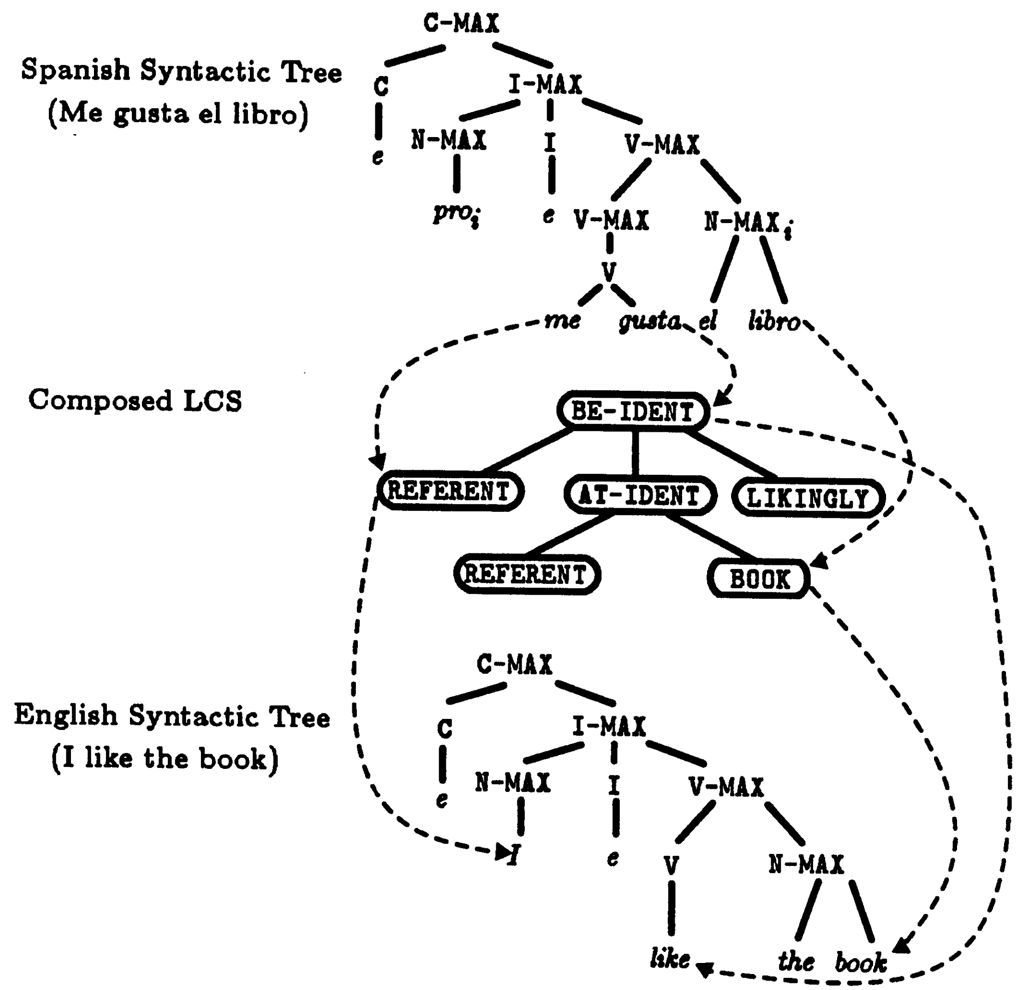

Figure 4: Translation of *Me gusta el libro* as *I like the book*

Notice that even though the arguments are not syntactically realized in the same way, the lexical selection procedure still succeeds. This is because of the separation between the syntactic description and the conceptual description. LCS descriptions provide the abstraction necessary for lexical selection without regard to syntax. In the next subsection, we will see how syntactic descriptions provide the necessary mechanism for argument realization without regard to conceptual considerations.

to the surface form. Pustejovsky and Nirenburg [1987] provide an elegant approach to generation of open-class lexical items based on focus information. Because the system described here does not include a model of discourse, the direct-mapping technique is used for such problematic cases.

3.2 Syntactic Realization: Structural Divergence and Conflation

Syntactic realization is the task of mapping a syntactic description to a surface-syntactic representation. Two problems are associated with this task. The first is that source- and target-language forms are potentially structurally divergent. An example of structural divergence is the realization of arguments in the translation of *tener* to *be* as in (5) (the corresponding argument-structures are included):

(5)
Yo tengo hambre $[_{\text{V-MAX}}\ [_{\text{V}}\ \text{tener}]\ [_{\text{N-MAX}}\ \text{hambre}]] \Rightarrow$
(I have hunger)
I am hungry $[_{\text{V-MAX}}\ [_{\text{V}}\ \text{be}]\ [_{\text{A-MAX}}\ \text{hungry}]]$

Here, not only are the predicates *tener* and *be* lexically distinct, but the arguments of these two predicates are structurally divergent: in Spanish, the argument is a noun-phrase, and, in English, the argument is an adjectival-phrase.

Figure 5 shows the LCS definitions used for this example. The equivalent LCS's for *tener* (1) and *be* provide the meaning "THING X is in an identificational state specified by PROPERTY Y." Note that there is another LCS for the word *tener* (2) that corresponds to a more literal translation (*have*) of the word *tener*.

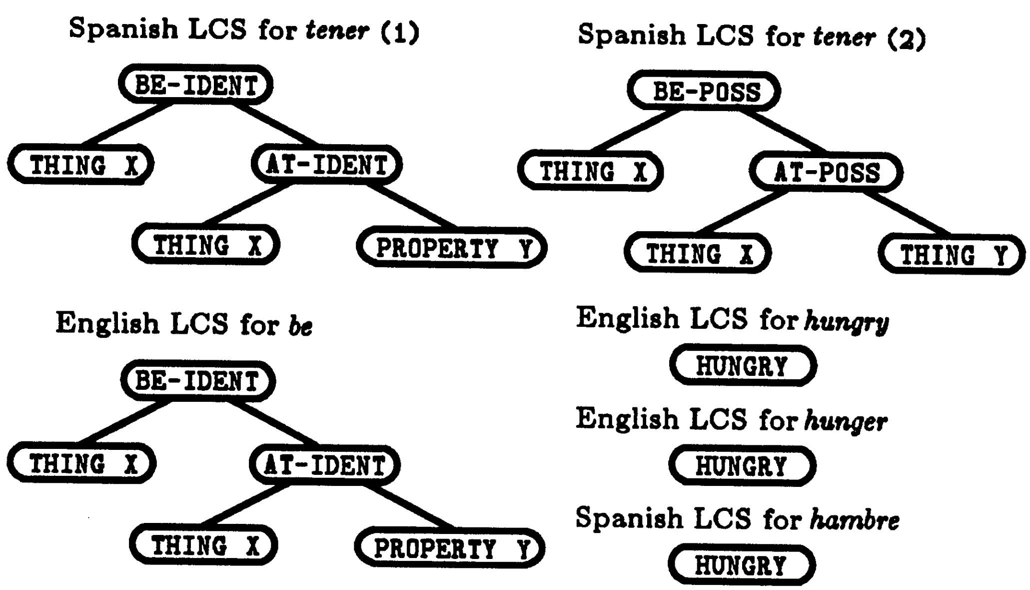

Figure 5: English and Spanish Lexical Entries for *tener-be* Example

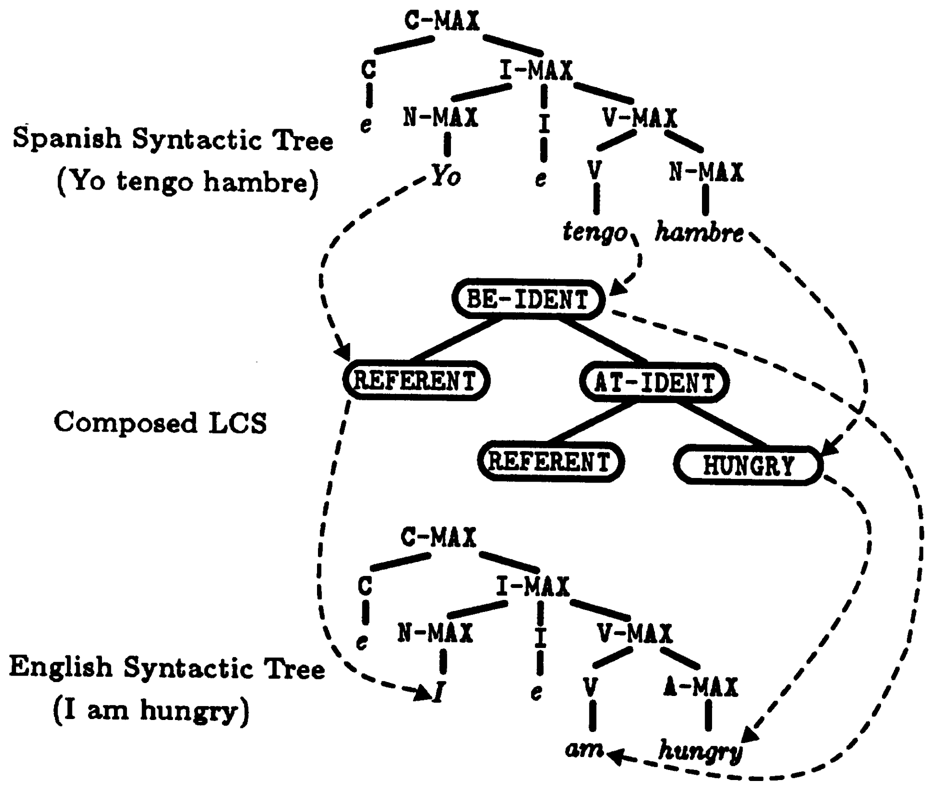

Figure 6: Translation of *Yo tengo hambre* as *I am hungry*

As for the lexical-selection of the appropriate predicate, the same LCS procedure that was used in the *stab-dar* case is used to match the LCS's of *tener* and *be*. However, for structural realization of the **PROPERTY** argument Y, the system must not only choose the appropriate lexical head, but it must choose the appropriate syntactic structure (*i.e.*, the category that will be projected from the head).

A syntactic-based scheme is inadequate for this example because it would choose the literal translation *hunger* for the source-language word *hambre*. This choice would be semantically awkward, but syntactically correct if the translation were *I have hunger*; however, if the more appropriate predicate *be* were chosen instead of *have*, the translation would be both semantically awkward and syntactically incorrect: *I am hunger*. A semantic-based scheme would make the correct lexical selection (that is, it would probably choose an argument that has a "desire to eat" property associated with it), but it would have no clue as to the syntactic form of the argument.

In the LCS approach, the lexical-selection procedure determines that both *hunger* and *hungry* lexically match the LCS for *hambre* because both are defined as the same LCS **HUNGRY** (which is a **PROPERTY**). In order to choose between these two possibilities, the system must access the `:INTERNAL` slot of the predicate *be* that was chosen as the top-level lexical head:

(6)
```
tener:  :INTERNAL ((Y N CONDITION))   :EXTERNAL ((X N AGENT))
be:     :INTERNAL ((Y A CONDITION))   :EXTERNAL ((X N AGENT))
```

Notice that, unlike the entry for *tener* the entry for *be* requires `Y` to be an adjective. Thus, the nominal possibility is eliminated, the adjective *hungry* is chosen, and the argument is projected up to its maximal level (`A-MAX`). Figure 6 shows the translation process for this example.

The second problem for structural realization is the potential for a divergent degree of *conflation* between the source- and target-language predicates. According to Talmy [1985], verbs may have a semantic representation that is not entirely exhibited at the level of syntactic structure. For example, the verb *enter* incorporates a *conflated* or "understood" particle *into* as part of its meaning structure; this particle manifests itself in the similar composite predicate *break into*. As it turns out, the Spanish equivalent of *break into* (*forzar*) has an additional conflated argument *entrada* (literally, *entry*); this argument is "understood," but not syntactically realized in English:

(7)
Juan forzó la entrada al cuarto (John forced entry to the room)
$[_{\text{V-MAX}}$ $[_{\text{V}}$ forzar$]$ $[_{\text{N-MAX}}$ la entrada$]$ $[_{\text{P-MAX}}$ a $\cdots]]$ $\Rightarrow$
John broke into the room
$[_{\text{V-MAX}}$ $[_{\text{V}}$ break$]$ $[_{\text{P-MAX}}$ into $\cdots]]$

Thus, there are three difficult tasks in the translation of *forzar* to *break*: selection of the predicate *break*, suppression (conflation) of the *entry* argument, and realization of the particle *into*.[10]

Figure 7 shows the LCS definitions used for this example. The LCS for *break* (1) provides the meaning "`THING X` goes locationally into `THING Y` forcefully." The LCS for *forzar* contains the `CAUSE` portion of this action, and the LCS for *entrada* contains the locational part of this action. Notice that the LCS definitions of *a* and *into* both have an `*EXTERNAL*` argument. The `*EXTERNAL*` marker is a place-holder for an LCS that will fill this position by means of lexical-semantic composition. For example, when the LCS associated with *a* is composed with the `GO-POSS` LCS, the argument that is the theme of the `GO-POSS` will replace the `*EXTERNAL*` marker of the *a* LCS.

A syntactic-based scheme has no notion of compositionality and would fail immediately in trying to map *forzar* (literally *force*) to *break* (or vice-versa). Furthermore, it would have the problem of choosing the appropriate particle,

[10]There are three analogous tasks in the reverse direction. That is, translation from English to Spanish requires selection of the predicate *forzar*, realization of the *entry* argument (this is actually an "inverse conflation"), and realization of the particle *a*.

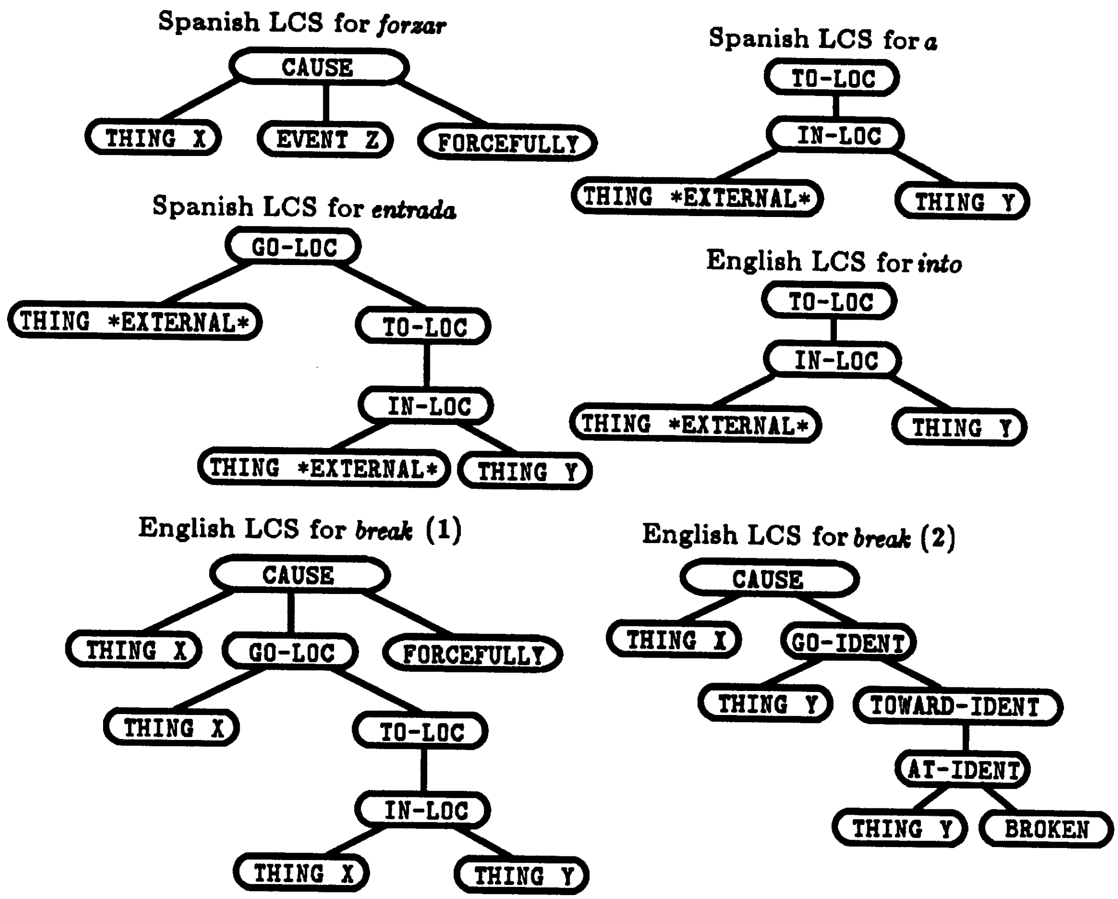

Figure 7: English and Spanish Lexical Entries for *forzar-break* Example

even if it were able to provide the correct structure (*i.e.*, a prepositional-phrase). On the other hand, a robust semantic-based scheme would have the ability to compose *forzar* and *entrada*, but it would not be able to determine whether the target-language argument was to be left implicit or whether it was to be syntactically realized, since there is no notion of conflation in such a scheme.

The LCS scheme uses compositionality to map *forzar la entrada* to *break*: the LCS for *forzar* contains a CAUSE, and the LCS for *entrada* contains a GO-LOC, both of which combine to match the composite LCS for *break*.

Notice that there are two LCS's for the word *break*; the first contains the matching GO-LOC LCS for this example, and the second one contains a GO-IDENT LCS that corresponds to "breaking an object." The mapping routine of the lexical selection procedure succeeds on the first one and (correctly) fails on the second one for the *break into* example.

At this point, the structural realization procedure determines that the internal **TO-IN PATH** argument of *break* is prepositional:

(8) break: :INTERNAL ((Y P GOAL LOCATION))
:EXTERNAL ((X N AGENT ANIMATE))

Since the internal argument must be prepositional, the system matches the TO-IN PATH with the TO-IN PATH LCS of *into*, and the phrase *into the room* is realized. Notice that the conflation task has been fulfilled: because the GO-LOC LCS is incorporated into the LCS definition for *break* (unlike the LCS definition for *forzar*), the English sentence does not syntactically realize this argument. Figure 8 shows the the translation process for this example.

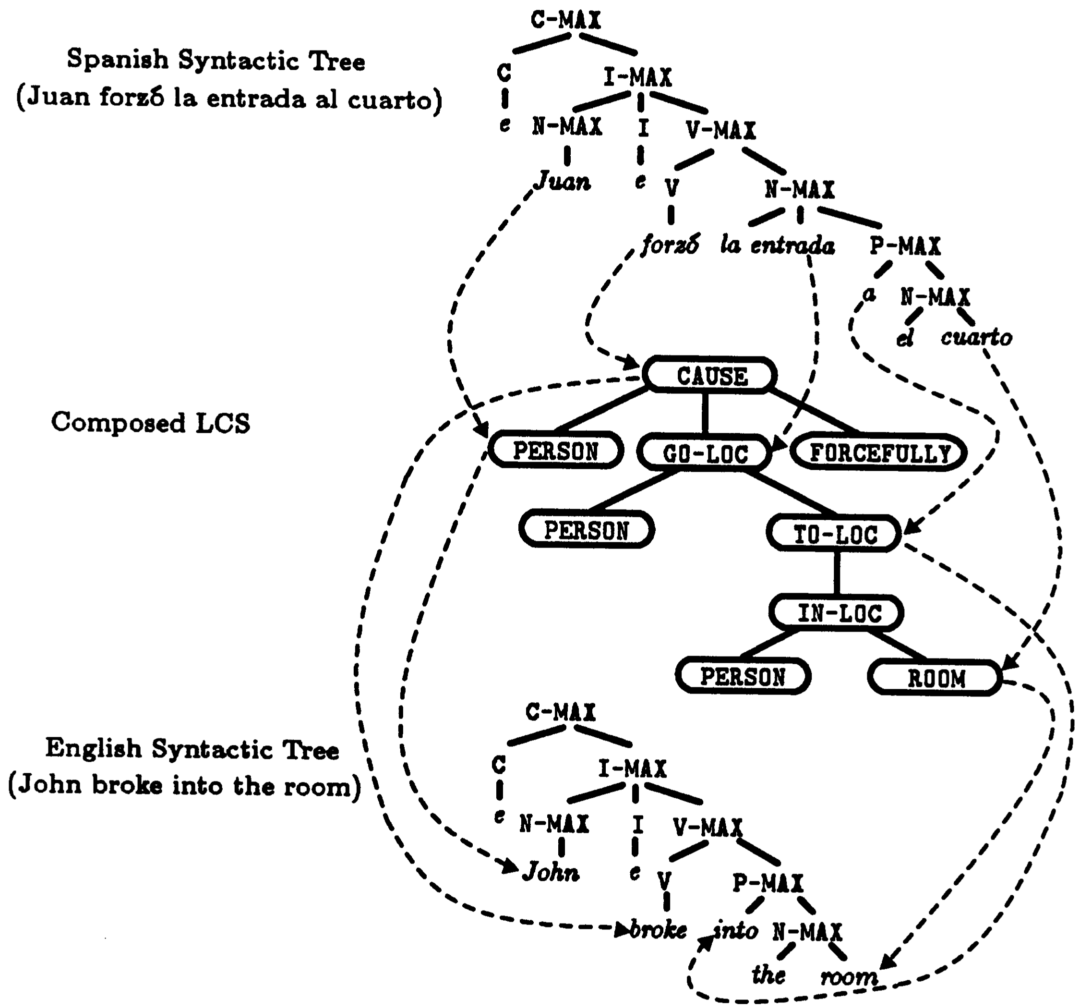

Figure 8: Translation of *Juan forzó la entrada al cuarto* as *John broke into the room*

4 Stab-Dar Revisited

We now return to our translation example: *I stabbed John.* Once the LCS for this sentence has been composed (see figure 2), the lexical selection procedure must choose the appropriate Spanish lexical head by matching the composed LCS not only at top level, but at all lower levels. Of the target-language root word possibilities that match the LCS **GO-POSS**, only the root word *dar* matches. Thus, this root is selected to be the lexical head that will be projected.

Next, the system must project the arguments of the selected lexical head *dar*. A recursive call is made to the selection procedure in order to determine the correct lexical head for each of the argument LCS's **REFERENT**, **KNIFE-WOUND** and **TOWARD**. Just prior to this recursive call, the system accesses the **:INTERNAL** and **:EXTERNAL** slots of the lexical head *dar* to establish the syntactic category that will be projected for each of these arguments. Notice that unlike the *stab* definition, the *dar* definition requires the **KNIFE-WOUND** to be realized as a noun phrase and the **TOWARD** argument to be realized as a prepositional phrase:

```
(9)  stab: :INTERNAL ((Y THEME KNIFE-WOUND) (Z N GOAL))
           :EXTERNAL ((X N AGENT ANIMATE))
     dar:  :INTERNAL ((Y N THEME KNIFE-WOUND) (Z P GOAL))
           :EXTERNAL ((X N AGENT ANIMATE))
```

Since the **KNIFE-WOUND** argument is not associated with a syntactic category in the English entry, it is not overtly realized, but conflated into the meaning of *stab*. Thus, the system performs an "inverse conflation" in order to arrive at the target-language realization for this example. The lexical heads chosen for LCS's **REFERENT**, **KNIFE-WOUND**, and **TOWARD** are *yo*, *cuchilladas*, and *a*, respectively. As dictated by the syntactic argument slots of the lexical head *dar*, these three heads are maximally-projected as **N-MAX**, **N-MAX**, and **P-MAX**, respectively. Finally, the **PERSON** LCS is projected as **N-MAX** according to the **:INTERNAL** slot of the lexical head *a*:[11]

```
(10) a: :INTERNAL ((Z N))
```

Figure 9 shows how the target-language tree is generated from the composed LCS.

[11]The proper noun *John* is considered to be a member of one of the many open-ended word classes discussed in footnote 9. Thus, the translation *Juan* is selected on the basis of a direct mapping from the source-language form.

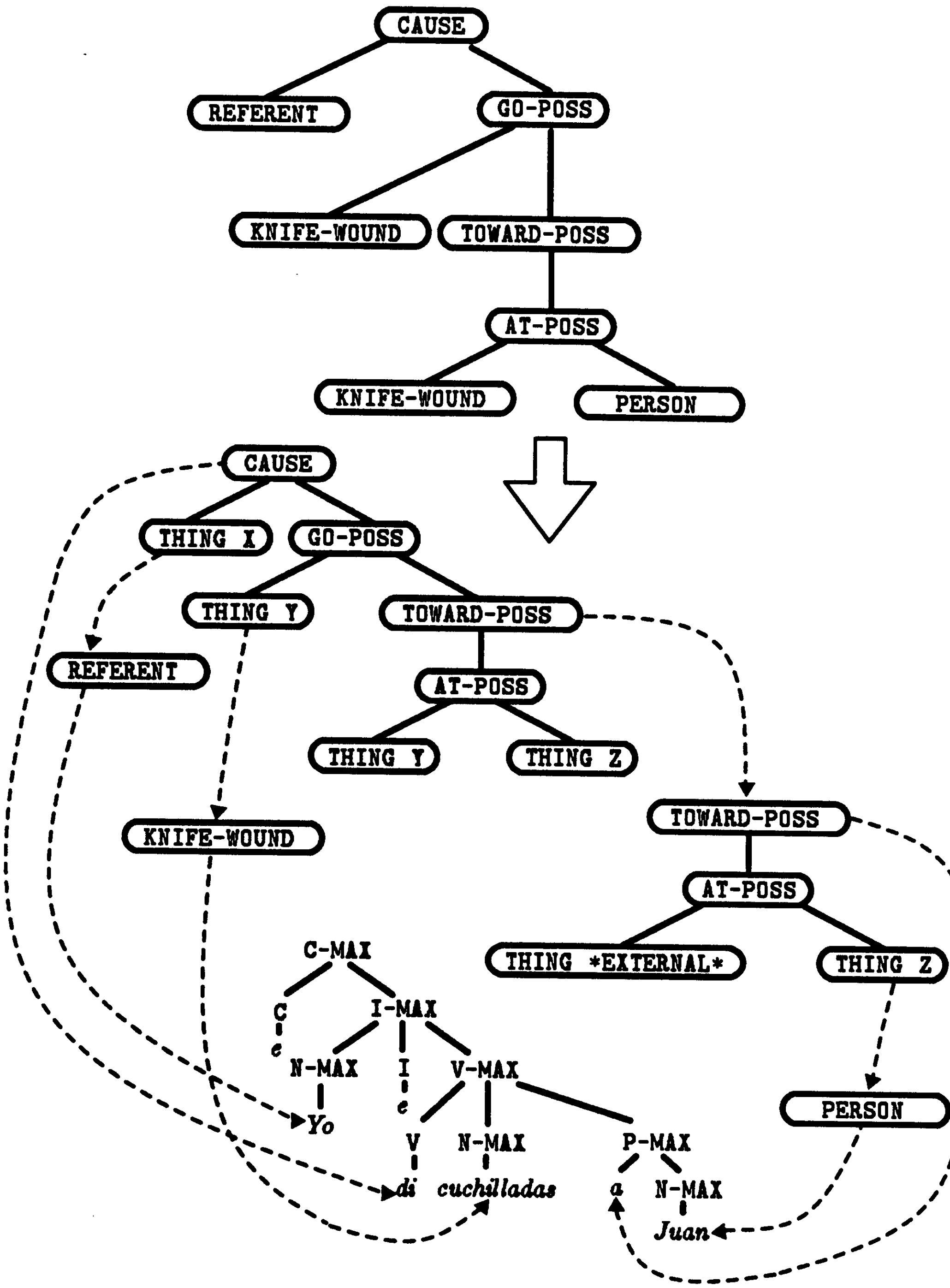

Figure 9: Generating from the Composed LCS for the *stab* Event

5 Summary

This report has demonstrated that lexical conceptual structure can be valuable for sentence generation, particularly in the context of machine translation. Two operations, lexical selection and syntactic realization, have been identified; in addition, two potential hazards, structural and thematic divergence, have been isolated. LCS descriptions seem to provide the abstraction necessary for selecting appropriate target-language terms with minimal dependence on syntax. In addition, LCS's provide the necessary mechanism for realizing arguments without regard to conceptual considerations. Although this approach is related to other generation approaches, it differs from syntactic-based approaches in that it avoids the non-compositional, direct-mapping word selection, and it differs from semantic-based approaches in that it does not entirely abandon syntactic considerations for word selection and structural realization. In summary, this report has shown that the combination of lexical-conceptual description and syntactic description facilitates the lexical-selection and structural realization processes, and it also aids in tackling the associated problems of thematic divergence, structural divergence, and conflation.

6 References

Cullingford, Richard E. (1986) *Natural Language Processing: A Knowledge-Engineering Approach*, Rowman and Littlefield, Totowa, New Jersey.

Dorr, Bonnie J. (1987) "UNITRAN: A Principle-Based Approach to Machine Translation," AI Technical Report 1000, Master of Science thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Dorr, Bonnie J. (1989) "Conceptual Basis of the Lexicon in Machine Translation," *First Annual Workshop on Lexical Acquisition, IJCAI-89*, Detroit, Michigan.

Hale, Kenneth and M. Laughren (1983) "Warlpiri Lexicon Project: Warlpiri Dictionary Entries," Massachusetts Institute of Technology, Cambridge, MA, Warlpiri Lexicon Project.

Hale, Kenneth and Jay Keyser (1986) "Some Transitivity Alternations in English," Center for Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA, Lexicon Project Working Papers #7.

Jackendoff, Ray S. (1983) *Semantics and Cognition*, MIT Press, Cambridge, MA.

Jackendoff, Ray S. (1989) *Semantic Structure*, unpublished manuscript, Brandeis University.

Lytinen, Steven L. (1987) "Integrating Syntax and Semantics," in *Machine Translation: Theoretical and Methodological Issues*, Sergei Nirenburg (ed.), Cambridge University Press, Cambridge, England.

McDonald, David D. (1987) "Natural Language Generation: Complexities and Techniques," in *Machine Translation: Theoretical and Methodological Issues*, Sergei Nirenburg (ed.), Cambridge University Press, Cambridge, England.

McKeown, Kathleen (1985) *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*, Cambridge University Press, Cambridge, England.

Nirenburg, Sergei, Victor Raskin, and Allen B. Tucker (1987) "The Structure of Interlingua in TRANSLATOR," in *Machine Translation: Theoretical and Methodological Issues*, Sergei Nirenburg (ed.), Cambridge University Press, Cambridge, England.

Pustejovsky, James and Sergei Nirenburg (1987) "Lexical Selection in the Process of Language Generation," *Proceedings of the 25th Annual Conference of the Association for Computational Linguistics*, Stanford University, Stanford, CA, 201–206.

Schank, Roger C. and Robert Abelson (1977) *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Slocum, Jonathan (1984) "METAL: The LRC Machine Translation System," Linguistics Research Center, University of Texas, Austin, Working Paper LRC-84-2.

Talmy, Leonard (1983) "How Language Structures Space," in *Spatial Orientation: Theory, Research, and Application*, Pick, Herbert L., Jr., and Linda P. Acredolo (eds.), Plenum Press, New York.