A PARAMETERIZED APPROACH TO INTEGRATING ASPECT WITH LEXICAL-SEMANTICS FOR MACHINE TRANSLATION

Bonnie J. Dorr*
Institute for Advanced Computer Studies
A.V. Williams Building
University of Maryland
College Park, MD 20742
bonnie@umiacs.umd.edu

ABSTRACT

This paper discusses how a two-level knowledge representation model for machine translation integrates aspectual information with lexical-semantic information by means of parameterization. The integration of aspect with lexical-semantics is especially critical in machine translation because of the lexical selection and aspectual realization processes that operate during the production of the target-language sentence: there are often a large number of lexical and aspectual possibilities to choose from in the production of a sentence from a lexical semantic representation. Aspectual information from the source-language sentence constrains the choice of target-language terms. In turn, the target-language terms limit the possibilities for generation of aspect. Thus, there is a two-way communication channel between the two processes. This paper will show that the selection/realization processes may be parameterized so that they operate uniformly across more than one language and it will describe how the parameter-based approach is currently being used as the basis for extraction of aspectual information from corpora.

INTRODUCTION

This paper discusses how the two-level knowledge representation model for machine translation presented by Dorr (1991) integrates aspectual information with lexical-semantic information by means of parameterization. The parameter-based approach borrows certain ideas from previous work such as the lexical-semantic model of Jackendoff (1983, 1990) and models of aspectual representation including Bach (1986), Comrie (1976), Dowty (1979), Mourelatos (1981), Passonneau (1988), Pustejovsky (1988, 1989, 1991), and Vendler (1967). However, unlike previous work, the current approach examines aspectual considerations within the context of machine translation. More recently, Bennett *et al.* (1990) have examined aspect and verb semantics within the context of machine translation in the spirit of Moens and Steedman (1988). This paper borrows from, and extends, these ideas by demonstrating how this theoretical framework might be adapted for cross-linguistic applicability. The framework has been tested within the context of an interlingual machine translation system and is currently being used as the basis for extraction of aspectual information from corpora.

(1) **Syntactic:**

(a) **Null Subject divergence:**
E: I have seen Mary ⇔ S: He visto a María
(Have seen (to) Mary)

(b) **Constituent Order divergence:**
E: I have seen Mary ⇔ G: Ich habe Marie gesehen
(I have Mary seen)

(2) **Lexical-Semantic:**

(a) **Thematic divergence:**
E: I like Mary ⇔ S: María me gusta a mí (Mary pleases me)

(b) **Structural divergence:**
E: John entered the house ⇔ S: Juan entró en la casa
(John entered in the house)

(c) **Categorial divergence:**
E: Yo tengo hambre ⇔ S: Ich habe Hunger (I have hunger)

(3) **Aspectual:**

(a) **Iterative Divergence:**
E: John stabbed Mary ⇔
S: Juan le dio una puñalada a María
(John gave a knife-wound to Mary)
S: Juan le dio puñaladas a María
(John gave knife-wounds to Mary)

(b) **Durative Divergence:**
E: John met/knew Mary ⇔
S: Juan conoció a María (John met Mary)
S: Juan conocía a María (John knew Mary)

Figure 1: Three Levels of MT Divergences

The integration of aspect with lexical-semantics is especially critical in machine translation because of the lexical selection and aspectual realization processes that operate during the production of the target-language sentence: there are often a large number of lexical and aspectual possibilities to choose from in the production of a sentence from a lexical semantic representation. Aspectual information from the source-language sentence constrains the choice of target-language terms. In turn, the target-language terms limit the possibilities for generation of aspect. Thus, there is a two-way communication channel between the two processes.

Figure 1 shows some of the types of parametric *divergences* (Dorr, 1990a) that can arise cross-linguistically.

*This paper describes research done in the Institute for Advanced Computer Studies at the University of Maryland. A special thanks goes to Terry Gaasterland and Ki Lee for helping to close the gap between properties of aspectual information and properties of lexical-semantic structure. In addition, useful guidance and commentary during this research were provided by Bruce Dawson, Michael Herweg, Jorge Lobo, Paola Merlo, Norbert Hornstein, Patrick Saint-Dizier, Clare Voss, and Amy Weinberg.

We will focus primarily on the third type, aspectual distinctions, and show how these may be discovered through the extraction of information in a monolingual corpus. We adopt the viewpoint that the algorithms for extraction of syntactic, lexical-semantic, and aspectual information must be well-grounded in linguistic theory. Once the information is extracted, it may then be used as the basis of parameterized machine translation. Note that we reject the commonly held assumption that the use of corpora necessarily suggests that statistical or example-based techniques be used as the basis for a machine translation system.

The following section discusses how the two levels of knowledge, aspectual and lexical-semantic, are used in an interlingual model of machine translation. We then describe how this information may be parameterized. Finally, we discuss how the automatic acquisition of new lexical entries from corpora is achieved within this framework.

TWO-LEVEL KR MODEL: ASPECTUAL AND LEXICAL-SEMANTIC KNOWLEDGE

The hypothesis proposed by Tenny (1987, 1989) is that the mapping between cognitive structure and syntactic structure is governed by aspectual properties. The implication is that lexical-semantic knowledge exists at a level that does not include aspectual information (though these two types of knowledge may depend on each other in some way). This hypothesis is consistent with the view adopted here: we assume that lexical semantic knowledge consists of such notions as predicate-argument structure, well-formedness conditions on predicate-argument structures, and procedures for lexical selection of surface-sentence tokens; all other types of knowledge must be represented at some other level.

Figure 2 shows the overall design of the UNITRAN machine translation system (Dorr, 1990a, 1990b). The system includes a two-level model of knowledge representation (KR) (see figure 2(a)) in the spirit of Dorr (1991). The translation example shown here illustrates the fact that the English sentence *John went to the store when Mary arrived* can be translated in two ways in Spanish. This example will be revisited later.

The lexical-semantic representation that is used as the interlingua for this system is an extended version of *lexical conceptual structure* (henceforth, LCS) (see Jackendoff (1983, 1990)). This representation is the basis for the lexical-semantic level that is included in the KR component. The second level that is included in this component is the aspectual structure.

The KR component is parameterized by means of *selection charts* and *coercion functions*. The notion of *selection charts* is described in detail in Dorr and Gaasterland (submitted) and will be discussed in the context of machine translation in the section on the Selection of Temporal Connectives. The notion of *coercion functions* was introduced for English verbs by Bennett *et al.* (1990). We extend this work by parameterizing the coercion functions and setting the parameters to cover Spanish; this will be discussed in the section on Selection and

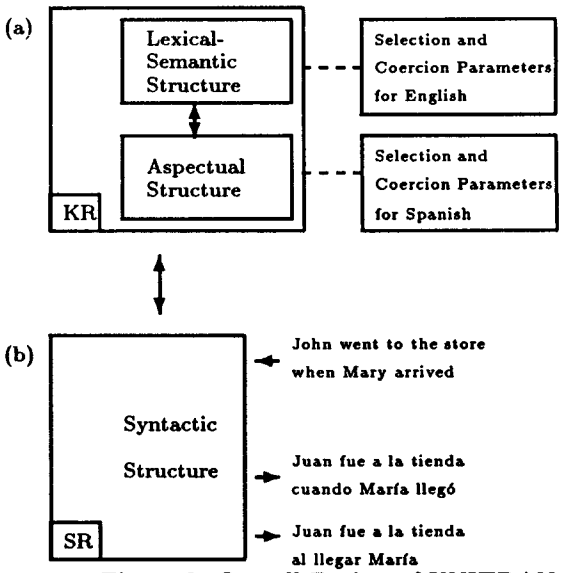

Figure 2: Overall Design of UNITRAN

Aspectual Realization of Verbs.

An example of the type of coercion that will be considered in this paper is the use of durative adverbials:

(4) (i) John ransacked the house { for an hour. / until Jack arrived. }

(ii) John destroyed the house { for an hour. / until Jack arrived. }

(iii) * John obliterated the house { for an hour. / until Jack arrived. }

Durative adverbials (*e.g.*, *for an hour* and *until ...*) are viewed as *anti-culminators* (following Bennett *et al.* (1990)) in that they change the main verb from an action that has a definite moment of completion to an action that has been stopped but not necessarily finished. For example, the verb *ransack* is allowed to be modified by a durative adverbial since it is inherently durative; thus, no coercion is necessary in order to use this verb in the durative sense. In contrast, the verb *destroy* is inherently non-durative, but it is *coerced* into a durative action by means of adverbial modification; this accounts for the acceptability of sentence (4)(ii).[1] The verb *obliterate* must necessarily be non-durative (*i.e.*, it is inherently non-durative and non-coercible), thus accounting for the ill-formedness of sentence (4)(iii).

In addition to the KR component, there is also a syntactic representation (SR) component (see figure 2(b)) that is used for manipulating the syntactic structure of a sentence. We will omit the discussion of the SR component of UNITRAN (see, for example, Dorr (1987)) and will concern ourselves only with the KR component for the purposes of this paper.

The remainder of this section defines the dividing line between lexical knowledge (*i.e.*, properties of predicates

[1] Some native speakers consider sentence (4)(ii) to be odd, at best. This is additional evidence for the existence of inherent features and suggests that, in some cases (*i.e.*, for some native speakers), the inherent features are considered to be absolute overrides, even in the presence of modifiers that might potentially change the aspectual features.

and their arguments) and non-lexical knowledge (*i.e.*, aspect), and discusses how these two types of knowledge are combined in the KR component.

Lexical-Semantic Structure. Lexical-semantic structure exists at a level of knowledge representation that is distinct from that of aspect in that it encodes information about predicates and their arguments, plus the potential realization possibilities in a given language. In terms of the representation proposed by Jackendoff (1983, 1990), the lexical-semantic structures for the two events of figure 2 would be the following:

(5) (i) [Event GO_{Loc} ([Thing John], [Position TO_{Loc} ([Thing John], [Location Store]))]

(ii) [Event GO_{Loc} ([Thing Mary], [Position TO_{Loc} ([Thing Mary], [Location e]))][2]

Although temporal connectives are not included in Jackendoff's theory, it is assumed that these two structures would be related by means of a lexical-semantic token corresponding to the temporal relation between the two events.

The lexical-semantic representation provided by Jackendoff distinguishes between events and states; however, this distinction alone is not sufficient for choosing among similar predicates that occur in different aspectual categories. In particular, events can be further subdivided into more specific types so that *non-culminative* events (*i.e.*, events that do not have a definite moment of completion) such as *ransack* can be distinguished from *culminative* events (*i.e.*, events that have a definite moment of completion) such as *obliterate*. This is a crucial distinction given that these two similar words cannot be used interchangeably in all contexts. Such distinctions are handled by augmenting the lexical-semantic framework so that it includes aspectual information, which we will describe in the next section.

Aspectual Structure. Aspect is taken to have two components, one comprised of inherent features (*i.e.*, those features that distinguish between states and events) and another comprised of non-inherent features (*i.e.*, those features that define the perspective, *e.g.*, simple, progressive, and perfective). This paper will focus primarily on inherent features.[3]

Previous representational frameworks have omitted aspectual distinctions among verbs, and have typically merged events under the single heading of *dynamic* (see, *e.g.*, Yip (1985)). However, a number of aspectually oriented lexical-semantic representations have been proposed that more readily accommodate the types of aspectual distinctions discussed here. The current work borrows extends these ideas for the development of an interlingual representation. For example, Dowty (1979) and Vendler (1967) have proposed a four-way aspectual classification system for verbs: states, activities, achievements, and accomplishments, each of which has a different degree of telicity (*i.e.*, culminated *vs.* nonculminated), and/or atomicity (*i.e.*, point *vs.* extended).[4] A similar scheme has been suggested by Bach (1986) and Pustejovsky (1989) (following Mourelatos (1981) and Comrie (1976)) in which actions are classified into states, processes, and events.

The lexical-semantic structure adopted for UNITRAN is an augmented form of Jackendoff's representation in which events are distinguished from states (as before), but events are further subdivided into activities, achievements, and accomplishments. The subdivision is achieved by means of three features proposed by Bennett *et al.* (1990) following the framework of Moens and Steedman (1988): ±dynamic (*i.e.*, events *vs.* states, as in the Jackendoff framework), ±telic (*i.e.*, culminative events (transitions) *vs.* nonculminative events (activities)), and ±atomic (*i.e.*, point events *vs.* extended events). We impose this system of features on top of the current lexical-semantic framework. For example, the lexical entry for all three verbs, *ransack, obliterate,* and *destroy,* would contain the following lexical-semantic representation:

(6) [Event CAUSE ([Thing X], [Event GO_{Loc} ([Thing X], [Position TO_{Loc} ([X John], [Property DESTROYED]))]))]

The three verbs would then be distinguished by annotating this representation with the aspectual features [+d,-t,-a] for the verb *ransack*, [+d,+t,-a] for the verb *destroy*, and [+d,+t,+a] for the verb *obliterate*, thus providing the appropriate distinction for cases such as (4).[5]

In the next section, we will see how the lexical-semantic representation and the aspectual structure are combined parametrically to provide the framework for generating a target-language surface form.

CROSS-LINGUISTIC APPLICABILITY: PARAMETERIZATION OF THE TWO-LEVEL MODEL

Although issues concerning lexical-semantics and aspect have been studied extensively, they have not been examined sufficiently in the context of machine translation. Machine translation provides an appropriate testbed for trying out theories of lexical semantics and aspect. The problem of lexical selection during generation of the target language is the most crucial issue in this regard. The current framework facilitates the selection of temporal connectives and the aspectual realization of verbs. We will discuss each of these, in turn,

[2]The empty location denoted by *e* corresponds to an unrealized argument of the predicate *arrive*.

[3]See Dorr and Gaasterland (submitted) for a discussion about non-inherent aspectual features.

[4]Dowty's version of this classification collapses achievements and accomplishments into a single event type called a *transition*, which covers both the point and extended versions of the event type. The rationale for this move is that all events have *some* duration, even in the case of so-called punctual events, depending on the granularity of time involved. (See Passonneau (1988) for an adaptation of this scheme as implemented in the PUNDIT system.) For the purposes of this discussion, we will maintain the distinction between achievements and accomplishments.

[5]This system identifies five distinct categories of predicates:

State:	[-d]	(like, know)
Activity (point):	[+d, -t, +a]	(tap, wink)
Activity (extended):	[+d, -t, -a]	(ransack, swim)
Achievement:	[+d, +t, +a]	(obliterate, kill)
Accomplishment:	[+d, +t, -a]	(destroy, arrive)

Matrix		Adjunct		Selected
Features	Perspective	Type	Perspective	Word
[±d,-t,±a]	perf	[+d,+t,±a]	simp, perf	When
[±d,-t,±a]	perfective	[+d,+t,±a]	simp, perf	Cuando
[±d,+t,±a]	perf	[+d,+t,±a]	simp, perf	Al

Figure 3: Selection Charts for *When*, *Cuando*, and *Al*

showing how *selection charts* and *coercion functions* are used as a means of parameterization for these processes.

Selection of Temporal Connectives: Selection Charts. In order to ensure that the framework presented here is cross-linguistically applicable, we must provide a mechanism for handling temporal connective selection in languages other than English. For the purposes of this discussion, we will examine distinctions between English and Spanish only.

Consider the following example:

(7) (i) John went to the store when Mary arrived.
(ii) John had gone to the store when Mary arrived.

In Dorr (1991), we discussed the selection of the lexical connective *when* on the basis of the temporal relation between the main or *matrix* clause and the subordinate or *adjunct* clause.[6] For the purposes of this paper, we will ignore the temporal component of word selection and will focus instead on how the process of word selection may be parameterized using the aspectual features described in the last section.

To translate (7)(i) and (ii) into Spanish, we must choose between the lexical tokens *cuando* and *al* in order to generate the equivalent temporal connective for the word *when*. In the case of (7)(i), there are two possible translations, one that uses the connective *cuando*, and one that uses the connective *al*:

(8) (i) Juan fue a la tienda cuando María llegó.
(ii) Juan fue a la tienda al llegar María.

Either one of these sentences is an acceptable translation for (7)(i). However, the same is not true of (7)(ii):[7]

(9) (i) Juan había ido a la tienda cuando María llegó.
(ii) Juan había ido a la tienda al llegar María.

Sentence (9)(i) is an acceptable translation of (7)(ii), but (9)(ii) does not mean the same thing as (7)(ii). This second sentence implies that John has already gone to the store and come back, which is not the preferred reading.

In order to establish an association between these connectives and the aspectual interpretation for the two events (*i.e.*, the matrix and adjunct clause), we compile a table, called a *selection chart*, for each language that specifies the contexts in which each connective may be used. Figure 3 shows the charts for *when*, *cuando*, and *al*.[8]

The selection charts can be viewed as inverted dictionary entries in that they map features to words, not words to features.[9] The charts serve as a means of parameterization for the program that generates sentences from the interlingual representation in that they are allowed to vary from language to language while the procedure for choosing temporal connectives applies cross-linguistically.[10] The key point to note is that the chart for the Spanish connective *al* is similar to that for the English connective *when* except that the word *al* requires the matrix event to have the +telic feature (*i.e.*, the matrix action must reach a culmination). This accounts for the distinction between *cuando* and *al* in sentences (9)(i) and (9)(ii) above.[11,12]

These tables are used for the selection of temporal connectives during the generation process (for which the relevant index into the tables would be the aspectual features associated with the interlingual representation). The selection of a temporal connective, then, is simply a table look-up procedure based on the aspectual features associated with the events.

Selection and Aspectual Realization of Verbs: Coercion Functions. Above, we considered the selection of temporal connectives without regard to the selection and aspectual realization of the lexical items that were being connected. Again, to ensure that the framework presented here is cross-linguistically applicable, we must provide a mechanism for handling lexical selection and aspectual realization in languages other than English.

Consider the English sentence *I stabbed Mary*. This may be realized in at least two ways in Spanish:[13]

(10) (i) Juan le dio puñaladas a María
(ii) Juan le dio una puñalada a María

[6]This work was based on theories of tense/time by Hornstein (1990) and Allen (1983, 1984).

[7]I am indebted to Jorge Lobo (personal communication, 1991) for pointing this out to me.

[8]The perfective and simple aspects are denoted as *perf* and *simp*, respectively.

[9]Note, however, that the features correspond to the events connected by the words, not to the words themselves.

[10]Because we are not discussing the realization of temporal information (*i.e.*, the time relations between the matrix and adjunct events), an abbreviated form of the actual chart is being used. Specifically, the chart shown in figure 3 assumes that the matrix event occurs *before* the adjunct event. See Dorr (1991) and Dorr and Gaasterland (submitted) for more details about the relationship between temporal information and aspectual information and the actual procedures that are used for the selection of temporal connectives.

[11]It has recently been pointed out by Michael Herweg (personal communication, 1991b) that the telic feature is not traditionally used to indicate a revoked consequence state (*e.g.*, the consequence state that results after returning from the "going to the store" event), but it is generally intended to indicate an irrevocable, culminative, consequence state. Thus, it has been suggested that *al* acts more as a complementizer than as a "pure" adverbial connective such as *cuando*; this would explain the realization of the adjunct not as a tensed adverbial clause, but as an infinitival subordinate clause. This possibility is currently under investigation.

[12]Space limitations do not permit the enumeration of the other selection charts for temporal connectives, but see Dorr and Gaasterland (submitted) for additional examples. Some of the connectives that have been compiled into tables are: *after, as soon as, at the moment that, before, between, during, since, so long as, until, while, etc.*

[13]Many other possibilities are available that are not listed here (*e.g.*, *Juan le acuchilló a María*).

Both of these sentences translate literally to "John gave stab wound(s) to Mary." However, the first sentence is the repetitive version of the action (*i.e.*, there were multiple stab wounds), whereas the second sentence is the non-repetitive version of the action (*i.e.*, there was only one stab wound). This distinction is characterized by means of the atomicity feature. In (10)(i), the event is associated with the features [+d,+t,-a], whereas, in (10)(ii) the event is associated with the features [+d,+t,+a].

According to Bennett *et al.* (1990), predicates are allowed to undergo an atomicity "coercion" in which an inherently non-atomic predicate (such as *dio*) may become atomic under certain conditions. These conditions are language-specific in nature, *i.e.*, they depend on the lexical-semantic structure of the predicate in question. Given the current featural scheme that is imposed on top of the lexical-semantic framework, it is easy to specify *coercion functions* for each language.

We have devised a set of coercion functions for Spanish analogous to those proposed for English by Bennett *et al.* The feature coercion parameters for Spanish differ from those for English. For example, the atomicity function does not have the same applicability in Spanish as it does for English. We saw this earlier in sentence (10), in which a singular NP verbal object maps a [-a] predicate into a [+a] predicate, *i.e.*, a non-atomic event becomes atomic if it is associated with a singular NP object. The parameterized mappings that we have constructed for Spanish are shown in figure 4(a). For the purposes of comparison, the analogous English functions proposed by Bennett *et al.* (1990) are shown in figure 4(b).[14]

Using the functions, we are able to apply the notion of feature-based coercion cross-linguistically, while still accounting for parametric distinctions. Thus, feature coercion provides a useful foundation for a model of interlingual machine translation.

A key point about the aspectual features and coercion functions is that they allow for a two-way communication channel between the two processes of lexical selection and aspectual realization.[15] To clarify this point, we return to our example that compares the three English verbs, *ransack*, *destroy*, and *obliterate* (see example (4) above). Recall that the primary distinguishing feature among these three verbs was the notion of telicity (*i.e.*, culminated *vs.* nonculminated). The lexical-semantic representation for all three verbs is identical, but the telicity feature differs in each case. The verb *ransack* is +telic, *obliterate* is -telic, and *destroy* is inherently -telic, although it may be coerced to +telic through the use of a durative adverbial phrase. Because *destroy* is a "co-

(a)

Spanish		
Mapping	*Parameters*	*Examples*
Telicity (C) f(-t)→+t	singular NP complements	Juan le dio una puñalada a María 'John stabbed Mary (once)'
	preterit past	Juan conoció a María 'John met Mary (once)'
Telicity (AC) f(+t)→-t	progressive morpheme	Lee estaba pintando un cuadro 'Lee was painting a picture (for some time)'
	imperfect past	Lee conocía a María 'Lee knew Mary (for some time)'
Atomicity f(+a)→-a	progressive morpheme	Chris está estornudando 'Chris is sneezing (repeatedly)'
	plural NP complements	Juan le dio puñaladas a María 'John stabbed Mary (repeatedly)'

(b)

English		
Mapping	*Parameters*	*Examples*
Telicity (C) f(-t)→+t	singular NP complements	John ran a mile
	culminative duratives	John ran until 6pm
Telicity (AC) f(+t)→-t	progressive morpheme	Lee was painting a picture
	non-culminative duratives	Lee painted the picture for an hour
Atomicity f(+a)→-a	progressive morpheme	Chris is sneezing
	frequency adverbials	Chris ate a sandwich everyday

Figure 4: Parameterization of Coercion Functions for English and Spanish

ercible" verb, it is stored in the lexicon as ±telic with a flag that forces -telic to be the inherent (*i.e.*, default) setting. Thus, if we are generating a surface sentence from an interlingual form that matches these three verbs but we know the value of the telic feature from the context of the source-language sentence (*i.e.*, we are able to determine whether the activity reached a definite point of completion), then we will choose *ransack*, if the setting is +telic, or *obliterate* or *destroy*, if the setting is -telic. In this latter case, only the word *destroy* will be selected if the interlingua includes a component that will be realized as a durative adverbial phrase.

Once the aspectual features have guided the lexical selection of the verbs, we are able to use these selections to guide the aspectual realizations that will be used in the surface form. For example, if we have chosen the word *obliterate* we would want to realize the verb in the simple past or present (*e.g.*, *obliterated* or *obliterate*) rather than in the progressive (*e.g.*, *was obliterating* or *is obliterating*). Thus, the aspectual features (and coercion functions) are used to choose lexical items, and the choice of lexical items is used to realize aspectual features.

The coercion functions are crucial for this two-way channel to operate properly. In particular, we must take care not to blindly forbid non-atomic verbs from being realized in the progressive since point activities, which are atomic (*e.g.*, *tap*), are frequently realized in the progressive (*e.g.*, *he was tapping the table*). In such cases the progressive morpheme is being used as an iterator of several identical atomic events as defined in the functions shown in figure 4. Thus, we allow "coercible" verbs

[14] Figure 4(b) contains a subset of the English functions. The reader is referred to Bennett *et al.* (1990) for additional functions. The abbreviations C and AC stand for culminator, and anti-culminator, respectively.

[15] Because the focus of this paper is on the lexical-semantic representation and associated aspectual parameters, the details of the algorithms behind the implementation of the two-way communication channel are not presented here; these are presented in Dorr and Gaasterland (submitted). We will illustrate the intuition here by means of example.

(*i.e.*, those that have a ±<feature> specification) to be selected and realized with the non-inherent feature setting if coercion is necessary for the aspectual realization of the verb.

ACQUISITION OF NOVEL LEXICAL ENTRIES: DISCOVERING THE LINK BETWEEN LCS AND ASPECT

In evaluating the parameterization framework proposed here, we will focus on one evaluation metric, namely the ease with which lexical entries may be automatically acquired from on-line resources. While testing the framework against this metric, a number of results have been obtained, including the discovery of a fundamental relationship between aspectual information and lexical-semantic information that provides a link between the primitives of Jackendoff's LCS representations and the features of the aspectual scheme described here.

Approach. A program has been developed for the automatic acquisition of novel lexical entries for machine translation.[16] We are in the process of building an English dictionary, and intend to use the same approach for building dictionaries in other languages, (*e.g.*, Spanish, German, Korean, and Arabic). The program automatically acquires aspectual representations from corpora (currently the Lancaster/Oslo-Bergen[17] (LOB) corpus) by examining the context in which all verbs occur and then dividing them into four groups: state, activity, accomplishment, and achievement. As we noted earlier, these four groups correspond to different combinations of aspectual features (*i.e.*, telic, atomic, and dynamic) that have been imposed on top of the lexical-semantic framework. Thus, if we are able to isolate these components of verb meaning, we will have made significant progress toward our ultimate goal of automatically acquiring full lexical-semantic representations of verb meaning.

The division of verbs into these four groups is based on several syntactic tests that are well-defined in the linguistic literature such as those by Dowty (1979) shown in figure 5.[18] Some tests of verb aspect shown here could not be implemented in the acquisition program because they require human interpretations. These tests are marked by asterisks (*). For example, Test 2 requires human interpretation to determine whether or not a verb has habitual interpretation in simple present tense.

The algorithm for determining the aspectual category of verbs is shown in figure 6. Note that step 3 applies Dowty's tests to a set of sentences corresponding to a particular verb until a unique category has been identified. In order for this step to succeed, we must ensure that Dowty's tests allow the four categories to be uniquely identified. However, a complication arises for the *state* category: out of the six tests that have been implemented from Dowty's table, only Test 1 uniquely sets states apart from the other three aspectual categories. That is, Test 1 is the only *implemented* test that has a value in the first column that is different from the other three columns. Note, however, that the value in this column is NO, which poses a problem for the above algorithm. Herein lies one of the major stumbling blocks for the extraction of information from corpora: it is only possible to derive new information in cases where there is a YES value in a given column. By definition, a corpus only provides *positive* evidence; it does not provide *negative* evidence. We cannot say anything about sentences that do *not* appear in the corpus. Just because a given sentence does not occur in a particular sample of English text does not mean that it can never show up in English. This means we are relying solely on the information that *does* appear in the corpus, *i.e.*, we are only able to learn something new about a verb when it corresponds to a YES in one of the rows of figure 5.[19]

Given that the identification of stative verbs could not be achieved by Dowty's tests alone, a number of hypotheses were made in order to identify states by other means. A preliminary analysis of the sentences in the corpus reveals that progressive verbs are generally preceded by verbs such as *be*, *like*, *hate*, *go*, *stop*, *start*, *etc.* These

	Test	STA	ACT	ACC	ACH
1.	X-ing is grammatical	no	yes	yes	yes
* 2.	has habitual interpretation in simple present tense	no	yes	yes	yes
3.	spend an hour X-ing, X for an hour	yes	yes	yes	no
4.	take an hour X-ing, X in an hour	no	no	yes	yes
* 5.	X for an hour entails X at all times in the hour	yes	yes	no	no
* 6.	Y is X-ing entails Y has X-ed	no	yes	no	no
7.	complement of stop	yes	yes	yes	no
8.	complement of finish	no	no	yes	no
* 9.	ambiguity with almost	no	no	yes	no
*10.	Y X-ed in an hour entails Y was X-ing during that hour	no	no	yes	no
11.	occurs with studiously, carefully, *etc.*	no	yes	yes	no

Figure 5: Dowty's Eleven Tests of Verb Aspect

1. Pick out main verbs from all sentences in the corpus and store them in a list called VERBS.
2. For each verb v in VERBS, find all sentences containing v and store them in an array SENTENCES[i] (where i is the indexical position of v in VERBS).
3. For each sentence set Sj in SENTENCE[j], loop through each sentence s in Sj:
   (a) Loop through each test t in figure 5.
   (b) See if t applies to s; if so, eliminate all aspectual categories with a NO in the row of figure 5 corresponding to test t.
   (c) Eliminate possibilities until a unique aspectual category is identified or until all sentences in SENTENCES have been exhausted.

Figure 6: Algorithm for Determining Aspectual Categories

[16]The implementation details of this program are reported in Dorr and Lee (1992).

[17]ICAME — Norwegian Computing Center for the Humanities (tagged version).

[18]This table is presented in Bennett *et al.* (1990), p. 250, based on Dowty (1979).

[19]Note that this is consistent with principles of recent models of language acquisition. For example, the *Subset Principle* proposed by Berwick (1985, p. 37) states that "the learner should hypothesize languages in such a way that positive evidence can refute an incorrect guess."

Verbs	Jackendoff Primitive	Aspectual Category	Aspectual Features
be	BE	state (STA)	[-d]
like	BE	state (STA)	[-d]
hate	BE	state (STA)	[-d]
go	GO	non-state (ACH)	[+d, +t, +a]
stop	GO	non-state (ACH)	[+d, +t, +a]
start	GO	non-state (ACH)	[+d, +t, +a]
finish	GO	non-state (ACH)	[+d, +t, +a]
avoid	STAY	non-state (ACT)	[+d, -t]
continue	STAY	non-state (ACT)	[+d, -t]
keep	STAY	non-state (ACT)	[+d, -t]

Figure 7: Circumstantial Verbs Categorized By Jackendoff's Primitives

Test to see if X appears in the progressive.

1. If YES, then apply one of the tests that distinguishes activities from achievements (*i.e.*, Test 3, Test 4, or Test 7).
2. If NO, apply Test 3 to rule out achievement or Test 4 to uniquely identify as an achievement.
3. Finally, if the aspectual category is not yet uniquely identified, either apply Test 11 to rule out activity or assume state.

Figure 8: Algorithm for Identifying Stative Verbs

verbs fall under a lexical-semantic category identified by Jackendoff (1983, 1990) as the circumstantial category. Based on this observation, the following hypothesis has been made:

Hypothesis 1: The only types of verbs that are allowed to precede progressive verbs are circumstantial verbs.

Circumstantial verbs subsume stative verbs, but they also include verbs in other categories. In terms of the lexical-semantic primitives proposed by Jackendoff (1983, 1990), the circumstantial verbs found in a subset of the corpus are categorized as shown in figure 7. An intriguing result of this categorization is that the circumstantial verbs provide a systematic partitioning of Dowty's aspectual categories (*i.e.*, states, activities, and achievements) into primitives of Jackendoff's system (*i.e.*, BE, STAY, and GO). Thus, the analysis of the corpora has provided a crucial link between the primitives of Jackendoff's LCS representation and the features of the aspectual scheme described earlier. If this is the case, then the framework has proven to be well-suited to the task of automatic construction of conceptual structures from corpora.

Assuming this partitioning is correct and complete, Hypothesis 1 can be refined as follows:

Hypothesis 1': The only types of verbs that are allowed to precede progressive verbs are states, achievements, and activities.

If this hypothesis is valid, the program is in a better position to identify stative verbs because it corresponds to a test that requires positive evidence rather than negative evidence. The hypothesis can be described by adding the following line to figure 5:

Test	STA	ACT	ACC	ACH
12. X <verb>-ing is grammatical	yes	yes	no	yes

Because there is a YES in the column headed by STA, verbs satisfying this test are potentially stative. Thus, once a verb X is found that satisfies this test, we apply the (heuristic) algorithm shown in figure 8 to determine whether X is stative.[20]

Verbs	Aspectual Category(s)
doing	(ACC)
facing	(ACC ACT)
asking	(ACC ACT)
made	(ACC)
drove	(ACC ACT)
welcome	(STA ACC ACT ACH)
emphasized	(STA ACC ACT ACH)
thanked	(ACC ACT STA)
staged	(ACC)
make	(ACC)
continue	(ACC ACT)
writes	(ACC)
building	(ACC)
running	(ACC ACT)
paint	(ACC)
finds	(ACC ACT)
arrives	(ACC ACT)
jailed	(ACC ACT STA)
nominating	(ACH ACT ACC)
read	(ACC ACT)
ensure	(STA ACC ACT ACH)
act	(ACT ACC)
carry	(ACC)
exercise	(ACC)
impose	(STA ACC ACT ACH)
contain	(STA ACC ACT ACH)
infuriate	(ACC ACT)

Figure 9: Aspectual Classification Results

Another hypothesis that has been adopted pertains to the distribution of progressives with respect to the verb *go*:

Hypothesis 2: The only types of progressive verbs that are allowed to follow the verb *go* are activities.

This hypothesis was adopted after it was discovered that constructions such as go running, go skiing, go swimming, *etc.* appeared in the corpus, but not constructions such as go eating, go writing, *etc.* The hypothesis can be described by adding the following line to figure 5:

Test	STA	ACT	ACC	ACH
13. go X-ing is grammatical	no	yes	no	no

The combination of Dowty's tests and these hypothesized tests allows the four aspectual categories to be more specifically identified.

Results and Future Work. Preliminary results have been obtained from running the program on 219 sentences of the LOB corpus (see figure 9).[21] Note that the program was not able to pare down the aspectual category to one in every case. We expect to have a significant improvement in the classification results once the sample size is increased.

Presumably more tests would be needed for additional improvements in results. For example, we have not proposed any tests that would guarantee the unique identification of accomplishments. Such tests are the subject of future research.

[20] Note that this algorithm does not guarantee that states will be correctly identified in all cases given that step 3 is a heuristic assumption. However, if Test 12 has applied, and state is still an active possibility, it is considerably safer to assume the verb is a state than it would be otherwise because we have eliminated accomplishments.

[21] For brevity, only a subset of the verbs are shown here.

In addition, research is currently underway to determine the restrictions (analogous to those shown in figure 5) that exist for other languages (*e.g.*, Spanish, German, Korean, and Arabic). Because the program is parametrically designed, it is expected to operate uniformly on corpora in other languages as well.

Another future area of research is the automatic acquisition of parameter settings for the construction of selection charts and aspectual coercion mappings on a per-language basis.

SUMMARY

This paper has examined a two-level knowledge representation model for machine translation that integrates aspectual information based on theories by Bach (1986), Comrie (1976), Dowty (1979), Mourelatos (1981), Passonneau (1988), Pustejovsky (1988, 1989, 1991), and Vendler (1967), and more recently by Bennett *et al.* (1990) and Moens and Steedman (1988), with lexical-semantic information based on Jackendoff (1983, 1990). We have examined the question of cross-linguistic applicability showing that the integration of aspect with lexical-semantics is especially critical in machine translation when there are a large number of temporal connectives and verbal selection/realization possibilities that may be generated from a lexical semantic representation. Furthermore, we have illustrated that the selection/realization processes may be parameterized, by means of selection charts and coercion functions, so that the processes may operate uniformly across more than one language. Finally, we have discussed the application of the theoretical foundations to the automatic acquisition of aspectual representations from corpora in order to augment the lexical-semantic representations that have already been created for a large number of verbs.

REFERENCES

Allen, James. F. (1983) "Maintaining Knowledge about Temporal Intervals," *Communications of the ACM* 26:11, 832–843.

Allen, James. F. (1984) "Towards a General Theory of Action and Time," *Artificial Intelligence* 23:2, 123–160.

Bach, Emmon (1986) "The Algebra of Events," *Linguistics and Philosophy* 9, 5–16.

Bennett, Winfield S., Tanya Herlick, Katherine Hoyt, Joseph Liro and Ana Santisteban (1990) "A Computational Model of Aspect and Verb Semantics," *Machine Translation* 4:4, 247–280.

Berwick, Robert C. (1985) *The Acquisition of Syntactic Knowledge*, MIT Press, Cambridge, MA.

Comrie, Bernard (1976) *Aspect*, Cambridge University Press, Cambridge, England.

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. (1990a) "Solving Thematic Divergences in Machine Translation," *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, University of Pittsburgh, Pittsburgh, PA, 127–134.

Dorr, Bonnie J. (1990b) "A Cross-Linguistic Approach to Machine Translation," *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Linguistics Research Center, The University of Texas, Austin, TX, 13–32.

Dorr, Bonnie J. (1991) "A Two-Level Knowledge Representation for Machine Translation: Lexical Semantics and Tense/Aspect," *Proceedings of the Lexical Semantics and Knowledge Representation Workshop, ACL-91*, University of California, Berkeley, CA, 250–263.

Dorr, Bonnie J. and Ki Lee (1992) "Building a Lexicon for Machine Translation: Use of Corpora for Aspectual Classification of Verbs," Institute for Advanced Computer Studies, University of Maryland, UMIACS TR 92-41, CS TR 2876.

Dorr, Bonnie J., and Terry Gaasterland (submitted) "Using Temporal and Aspectual Knowledge to Generate Event Combinations from a Temporal Database," *Third International Conference on Principles of Knowledge Representation and Reasoning*, Cambridge, MA, 1992.

Dowty, David (1979) *Word Meaning and Montague Grammar*, Reidel, Dordrecht, Netherlands.

Herweg, Michael (1991a) "Aspectual Requirements of Temporal Connectives: Evidence for a Two-level Approach to Semantics," *Proceedings of the Lexical Semantics and Knowledge Representation Workshop, ACL-91*, University of California, Berkeley, CA, 152–164.

Hornstein, Norbert (1990) *As Time Goes By*, MIT Press, Cambridge, MA.

ICAME — Norwegian Computing Center for the Humanities (tagged version) *Lancaster/Oslo-Bergen Corpus*, Bergen University, Norway.

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

Jackendoff, Ray S. (1990) *Semantic Structures*, MIT Press, Cambridge, MA.

Lobo, Jorge (1991) personal communication.

Moens, Marc and Mark Steedman (1988) "Temporal Ontology and Temporal Reference," *Computational Linguistics* 14:2, 15–28.

Mourelatos, Alexander (1981) "Events, Processes and States," in *Tense and Aspect*, P. J. Tedeschi and A. Zaenen (eds.), Academic Press, New York, NY.

Passonneau, Rebecca J. (1988) "A Computational Model of the Semantics of Tense and Aspect," *Computational Linguistics* 14:2, 44–60.

Pustejovsky, James (1988) "The Geometry of Events," Center for Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA, Lexicon Project Working Papers #24.

Pustejovsky, James (1989) "The Semantic Representation of Lexical Knowledge," *Proceedings of the First Annual Workshop on Lexical Acquisition, IJCAI-89*, Detroit, Michigan.

Pustejovsky, James (1991) "The Syntax of Event Structure," *Cognition*.

Tenny, Carol (1987) "Grammaticalizing Aspect and Affectedness," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Tenny, Carol (1989) "The Aspectual Interface Hypothesis," Center for Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA, Lexicon Project Working Papers #31.

Vendler, Zeno (1967) "Verbs and Times," *Linguistics in Philosophy*, 97–121.

Yip, Kenneth M. (1985) "Tense, Aspect and the Cognitive Representation of Time," *Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics*, Chicago, IL, 18–26.