A Multi-Level Approach to Interlingual MT: Defining the Interface between Representational Languages

Bonnie J. Dorr and Clare R. Voss

Department of Computer Science
University of Maryland
College Park, MD 20742
{bonnie,voss}@cs.umd.edu

Abstract

This paper describes a multi-level design, i.e., a non-uniform approach to interlingual machine translation (MT), in which distinct representational languages are used for different types of knowledge. We demonstrate that a linguistically-motivated "division of labor" across multiple representation levels has not complicated, but rather has readily facilitated, the identification and construction of systematic relations at the interface between each level. Our approach assumes an interlingua derived from the lexical semantics and predicate decomposition approaches of Jackendoff (1983; 1990) and Levin and Rappaport-Hovav (1995a; 1995b). We describe a model of interpretation and representation of natural language sentences which has been implemented as part of an interlingual MT system called PRINCITRAN.

1 Introduction

In order to produce an appropriate translation of a source-language sentence, a machine translation (MT) system must have access to several different representation types. Some examples of these are: *lexical*, for lexicon-based information; *syntactic*, for defining phrase structure; *interlingual* (or IL) for sentence interpretation; and *knowledge representational* (or KR) for filtering out interpretations that are incompatible with facts in the MT system's knowledge base. This paper examines the interface between the interlingua and other representation types in an interlingual MT system. We adopt a multi-level design (Dorr and Voss, 1993), i.e., a non-uniform approach, in which distinct representational languages are used for different types of knowledge. In our research and MT system construction, we have found that a linguistically-motivated "division of labor" in the translation task across multiple representation levels has not complicated, but rather has readily facilitated, the identification and construction of systematic relations at the interface between each level. We argue that our specific NLP application, the task of MT, is not hindered by a non-uniform approach.

Figure 1 illustrates the multi-level approach to interlingual MT. We adopt three different representational levels—syntactic, IL, and KR—each corresponding to a different processing phase in the system: (i) an analysis/synthesis phase in which a source-language (SL) sentence is parsed into a syntactic structure or a target-language (TL) sentence is generated from a syntactic structure and associated lexical items; (ii) a composition/decomposition phase in which a SL syntactic structure is composed into an IL representation or an IL representation is decomposed into a TL syntactic structure and lexical items; and (iii) a KR phase that

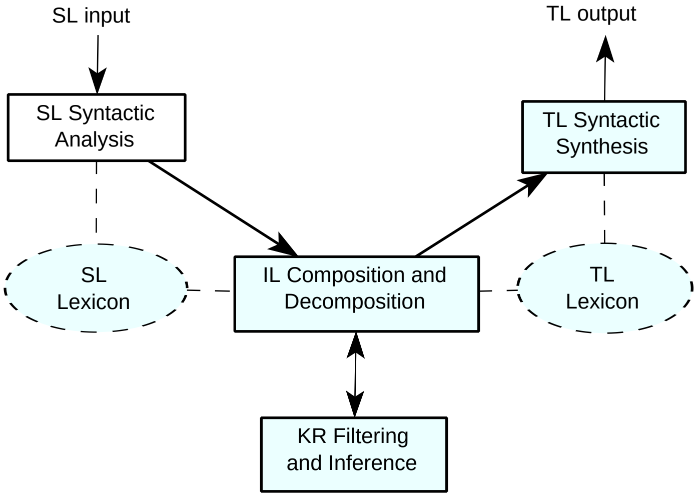

Figure 1: Multi-Level Approach to Interlingual MT: Processing Phases

checks the IL representations, filtering out those forms incompatible with known facts and, as needed, coercing or augmenting IL forms with logically inferred knowledge in order to resolve an incomplete IL composition. The model assumes that the SL and TL lexicons are each accessed, respectively, during the composition and decomposition of the IL form.[1] Figure 2 illustrates, for one sense of the intransitive verb *stand,* the static forms stored at the syntactic and IL levels and the primitives from within the IL form that are concepts grounded at the KR level.

As argued by Dorr and Voss (1993), the field of MT research lacks a consensus on what an *interlingua* is and how it is defined. MT system developers in building their individual interlinguas have selected from among a variety of semantic formalisms as the basis for their IL. For example, the Rosetta system (Rosetta, 1994) used an interlingua based on M-grammar, a representation derived from Montague Grammar. In the MT system Mikrokosmos, Levin and Nirenburg (1994) have been developing an interlingua based on their own Text Meaning Representation (TMR) language. Rupp, Johnson, and Rosner (1992) have worked with Situation Schemata, inspired by the Situation Semantics of Barwise and Perry (1983), for their semantic representation language. Recently Mani (1995), following insights from Zwarts and Verkuyl (1994), has proposed a "layered" interlingua whose forms contain Discourse Representation Structures (DRSs of Kamp (1981)) at one level and Lexical Conceptual Structures (LCSs of Jackendoff (1983; 1990)) at another level of representation.

Our approach assumes an interlingua derived from the lexical semantics and predicate decomposition approaches of Jackendoff (1983; 1990) and Levin and Rappaport-Hovav (1995a; 1995b). We describe a model of interpretation and representation of natural language sentences which has been implemented as part of an interlingual MT system called PRINCITRAN. This system combines the syntactic processing design of PRINCIPAR (Dorr, Lin, Lee, and Suh (1995)) with the syntax-IL interface originally developed in the UNITRAN

[1] Throughout this paper *SL* and *TL lexicons* will refer to the IL forms within the natural language-specific lexicons of the MT system. These lexicons are also accessed for their syntactic structures. For further details about the syntactic processing mechanism, see Dorr, Lin, Lee, and Suh (1995).

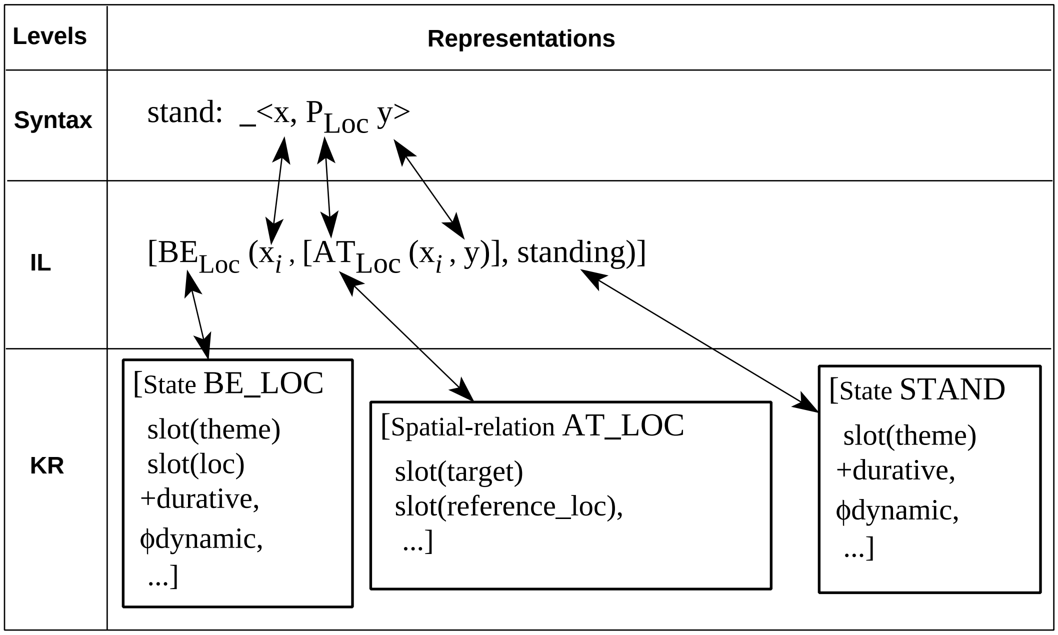
Levels
Representations
Syntax
stand: _<x, P_Loc y>
IL
[BE_Loc (x_i, [AT_Loc (x_i, y)], standing)]
KR
[State BE_LOC slot(theme) slot(loc) +durative, ϕdynamic, ...]
[Spatial-relation AT_LOC slot(target) slot(reference_loc), ...]
[State STAND slot(theme) +durative, ϕdynamic, ...]

Figure 2: Multi-Level Approach to Interlingual MT: Static Forms

system (Dorr, 1993) and the IL-KR interface from the LEXITRAN system (Dorr and Voss, 1993; Dorr et al., 1994).

Throughout this paper we aim to demonstrate the utility of a linguistically-motivated division into distinct representational languages in a multi-level IL MT design. The next section describes our framework for interlingual MT, citing examples from the spatial domain in support of our non-uniform representational approach. Section 3 describes our framework for defining an interlingual representation. The next two sections describe how the basic units of the IL representation serve as the mediating structures at the syntax/IL interface (Section 4) and at the IL/KR interface (Section 5).

2 Framework for Interlingual MT

We adopt a non-uniform approach to MT in that knowledge is represented at different levels using distinct representational formalisms. First, we give an example of translation data that led us to explore this non-uniform or "division of labor" approach. We then spell out the assumptions inherent in our MT system design. Next, we present the problem space and the subset of natural language that we have been working with. Finally, we describe our modularization of knowledge encodings into different MT components.

2.1 Motivation for Non-Uniform Approach

The separation of knowledge into different representational levels is motivated by cases such as the following German example:

(1) Die Kirche liegt im Süden der Stadt

which may have either of the following interpretations:

(2) (i) The church lies in the south of the city

(ii) The church lies to the south of the city

In the first interpretation, the *church* is located in the southern part of the city (i.e., within city limits), whereas in the second case, the *church* lies south of the city (i.e., outside its limits).

In (1), the ambiguity is not readily noticeable in the words of the German sentence, yet the conceptual distinction underlying the ambiguity (i.e., lying *inside of*, vs. lying *outside of*) is strikingly clear. That is, this translation data enables us to see that the German phrase *im Süden der Stadt* maps to two distinct representations: SOUTH-AND-INTERNAL corresponding to the southern region of the city and SOUTH-AND-EXTERNAL corresponding to the region south of the city.[2]

If (1) referred to a mountain rather than a church, the MT system should be able to use default knowledge in the KR that mountains are physical entities that are typically distinct and external to cities (thus choosing the second translation); yet, the system should also take advantage of specific facts in the KR, e.g., that a particular mountain is in the city, in order to override default knowledge as needed (thus choosing the first translation). We take this to be a KR filtering function that is independent of particular lexical knowledge. The need to translate such sentences accurately presents a clear case of where general as well as specific real world knowledge should assist in eliminating inappropriate translations. In the model we adopt, it is the KR, not the IL, that provides this capability.

The non-uniform approach alleviates many of the problems that might be encountered if one were to adopt an MT system design with only one representational language to assign interpretations to natural language input. In particular, the separation of the IL from the KR allows us to build a MT system in which we need not represent the "full" meaning for each word in a sentence being translated.[3] Furthermore, with the multi-level approach we can let the choice of interlingual structures be driven by the limited demands of the syntax-to-IL mapping, rather than by the full complexity of conceptual categories for events and states that properly belong in the KR system. This design consideration protects our system from becoming unnecessarily brittle as the KR system grows or changes with the domain of translation. It also reflects our bias toward maintaining the advantages of assumptions made by Dorr (1993) over those of Nirenburg et al. (1992) in IL-based MT system design when the two have different consequences for PRINCITRAN.

[2] Note that the French translation *l'église est au sud de la ville* has the same ambiguity. For readers not versed in German or French, consider the ambiguity in English: *he drove to the south of the city* has distinct readings for SOUTH-AND-INTERNAL corresponding to the southern region of the city and SOUTH-AND-EXTERNAL corresponding to the region south of the city. We will not address here the problem of *lexical allocation,* i.e., determining whether to allocate the ambiguity lexically to *Süden* or *in dem* in the phrase.

[3] Thus, we keep the fine-grained semantic distinctions that are specific to particular languages and poorly motivated by conceptual categories, separate from the IL level of representation. We hold the view that a mapping from *all* possible lexical semantic distinctions, fine-grained and otherwise, to the conceptual categories of the KR amounts to a strong version of the Whorfian hypothesis and, thus, we reject the notion of such a mapping.

2.2 Assumptions

We assume processing to be sentence-level only (i.e., no discourse analysis) and we take the output of the PRINCIPAR system (the parser used in PRINCITRAN as described by Dorr, Lin, Lee, and Suh (1995)) as our input. A SL sentence is analyzed into a *set* of parse trees representing all possible syntactic interpretations, within the Government-Binding theory of syntax upon which our parser is based. It is then the job of IL and KR components to interpret the SL phrase structure and provide an appropriate interlingua for the generation of the TL. Thus, there are two types of interfaces that we define between our three knowledge levels: (i) one that relates IL representations to corresponding syntactic forms by means of lexical entries; and (ii) one that checks the IL representations in the KR, filtering out those forms incompatible with known facts and, as needed, coercing or augmenting IL forms with logically inferred knowledge in order to resolve an incomplete IL composition. (See figure 1.)

The IL forms themselves are defined in terms of those two types of interfaces. As can be seen in the example of a lexical IL form in figure 2, the IL forms retain linguistically-relevant information from both interfaced levels: (i) structural components of lexical knowledge such as predicate-argument relations (e.g., that spatial verbs such as *stand, lie* involve an ordered binary relation on a located object and a location) and (ii) conceptual content such as predicate primitives and prototypical constants (e.g., that spatial zero-related verbs[4] such as *shovel* in *she shoveled the snow from the driveway* include generic event-type concepts such as REMOVAL activities and object-type concepts such as SHOVEL). Thus the IL serves as an interfacing level that mediates between the KR and its corresponding syntactic realization.

2.3 Defining the Problem Space: Translation Mismatches

We seek to address issues concerning translation mismatches. This is a problem area that has received increasingly greater attention in recent literature (see (Dorr and Voss, 1993; Barnett et al., 1994; Beaven, 1992; Kameyama et al., 1991; Kinoshita, Phillips, and Tsujii, 1992; Lindop and Tsujii, 1991; Whitelock, 1992) as well as related discussion in (Melby, 1986; Nirenburg and Nirenburg, 1988)). In particular, Barnett et al. (1994) identify two categories of "differences" between the source language and the target language: translation *divergences*, in which the same content is overtly conveyed in the source and target texts, but the structures of the sentences are different (as also defined in previous work by Dorr (1994)); and translation *mismatches*, in which the content that is conveyed is different in the source and target languages (as in (Kameyama et al., 1991)). Both types of distinctions must be addressed in translation, yet most MT researchers have attended to one or the other.

Researchers investigating *divergences* (see, e.g., Dorr and Voss (1993)) are more inclined to address the mechanism that links the IL representation to the syntactic structure of the target language, whereas investigators of the *mismatch* problem (see, e.g., (Barnett et al., 1994; Kameyama et al., 1991; Wu and Palmer, 1994; Palmer and Wu, 1995)) are more inclined to focus on the details of the conceptual representation underlying the IL. The novelty of our approach is that it addresses the problem of mismatches through access to the KR while retaining enough of the structure in the IL to resolve the divergence problem.

[4] We take the term "zero-related verb" from Levin (1993) for verbs either derived from nouns or out of which nouns have been derived.

We have examined the problem space within the domain of spatial predicates; we view the range of lexical mismatches in this domain as a set of equivalence classes. Figure 3 shows a sample of lexical mismatch classes.[5] This partitioning is designed to provide a framework within which a solution for one example in a partition or problem class will work on all the examples in that class. Operationally in IL-based MT, the "mismatches" occur during generation, when matching an IL form in the TL lexicon, but finding no TL lexical entry (a gap in lexicalization), or when matching on more than one TL entry (by lexical synonymy or differential lexicalizations). We do not attempt here to cover the full space of lexical mismatches; rather, we restrict our attention to a few of those that occur at the syntax-IL and IL-KR interface. In particular, we have left aside those that must be resolved by reference to broader contextual knowledge of transactions.[6]

- gaps in lexicalization for zero-related verbs[7]
  - no English word for *go by vehicle*
    but *bus, train, jet* (go by bus, train, jet)
  - no German word for *go*
    but *gehen, fahren* (go on foot, go by vehicle)
- lexical synonymy
  - English *lift, raise*
  - French *lever, hausser*
- differential lexicalization of caused/non-caused events
  - English non-causative *fall* and causative *fell/make fall*
    but non-causative and causative *break*
  - French non-causative *tomber* and causative *faire tomber*
    but non-causative and causative *casser*

Figure 3: A Sample of Lexical Mismatch Classes within the Spatial Domain

[5] Some of these examples are taken from related work by DiMarco, Hirst, and Stede (1993). Their research addresses many of the same questions we are examining.

[6] To give the reader a sense of what these KR level mismatches entail, consider the case of how the British, Swiss French and Swiss Germans lexicalize the same transaction during a bus ride: *punch the ticket, validate the ticket,* and *invalidate the ticket* (Kay, Gawron, and Norvig, 1994). Each focuses on a different aspect of the overall transaction, one on the action itself (the punching), one on the state of the ticket during the ride (a valid ticket), and one on the state of the ticket after the ride (an invalid ticket). These are more properly labeled *transaction-focus* KR mismatches.

[7] A reviewer points out that many vehicle names do not have zero-related verbs, e.g., *car*, *truck*, and *fiacre*, etc. We are not assuming that all vehicles names have a verbal counterpart; rather, our intention is to cover the cases where a zero-related verbal counterpart *does* exist. Note that the class of such verbs is larger than one might expect: *balloon, bicycle, bike, boat, bobsled, bus, cab, canoe, caravan, chariot, coach, cycle, dogsled, ferry, gondola, helicopter, jeep, jet, kayak, moped, motor, motorbike, motorcycle, parachute, punt,*

At generation time, we make the assumption that, for each SL word, there exists at least one TL word that is closest in meaning. From this it follows that, when an exact TL word match is missing (a gap), there are three possible relations between the closest TL word and the SL word: subsumes, subsumed-by, and overlapping, i.e., the closest TL word is *over-general*, *over-specific*, or *overlapping* in meaning with respect to SL word's meaning.[8] Below we will examine examples in each of these classes and show that in our approach, although the TL word that is initially selected does not exactly match the meaning of the corresponding SL word, a *full-coverage* meaning match is ultimately obtained by ensuring that some *combination* of TL words matches the overall SL concept.[9]

2.4 Domain of Inquiry: Spatial Predicates

The primary focus of our investigation is in the domain of simple spatial expressions, with specific attention to spatial verbs and prepositions that, collectively at the IL level, we label *spatial predicates*. Such predicates are used to describe relations between physical objects in real 3-dimensional space (e.g., a cup on a table). We take this to be a critical area of inquiry for interlingual MT since this is where a very high level of cross-linguistic diversity has been shown to exist. Consider the following English/French translations:[10]

(3) (i) English: The car roared down the street
(ii) French: La voiture descendit la rue en vrombissant
'The car went down the street (in) roaring'

(4) (i) English: The truck rumbled into the yard
(ii) French: Le camion entra dans la cour dans un grand fracas
'The truck entered in the yard in a big din'

English permits spatial directional predicates in phrases such as *down the street* and *into the yard* to compose readily with verbs of sound emission such as *roar* and *rumble.* In such constructions, we understand that the car and truck are moving as they make these noises even though there are no words that overtly refer to the motion. French, by contrast, does not permit such compositions: the motion and sound emission each must be lexicalized separately, as two verbs (e.g., in (3)(ii)) or as a verb and an extra modifier phrase in (e.g., in (4)(ii)). Thus, there is a mismatch in translating from English into French. Compared to English, French has a gap in its lexicalization space: French has no word equivalent to the *roar* and *rumble* in sentences above, meaning *go with a roaring/rumbling noise*. Consequently, in the analysis phase of translations, the MT system must identify the motion implicit in these English sentences, encode this in the IL representation, and then in

raft, rickshaw, rocket, ship, skate, skateboard, ski, sled, sledge, sleigh, taxi, toboggan, tram, trolley, yacht, etc.

[8] It has been argued in fuller detail elsewhere that these three relations span the full range of possible meaning mismatches (Barnett et al., 1994).

[9] By allowing a *combination* of TL words to be generated, we account for compounds such as *lift up*, *climb down*, etc.

[10] These examples are taken from (Levin and Rappaport-Hovav, 1995a). For other examples, also see the extensive discussions in (Talmy, 1983; Talmy, 1985).

the generation phase, make the motion lexically explicit when the TL is French.[11] Recent linguistic research indicates that Hebrew patterns like English in this regard while Japanese patterns like French.

In addition to such sentential structural distinctions found cross-linguistically, there also exists wide variation in the language-specific canonical surface location of spatial relations among the lexemes and morphemes of a sentence. For example, the spatial relation expressed in a preposition in English may appear in the combination of a verbal prefix and an overt preposition in a Russian translation, or as a postfix on the head noun of an NP in a language such as Korean. Or the equivalent of the English preposition may not actually appear as an overt distinct surface element in a Spanish translation, but instead be incorporated into the meaning of a verb.

One last note concerning this domain. While NL predicates in the spatial domain may need, for operational reasons, to be reduced to mathematical relations (e.g., described by Cartesian coordinates or topological models) for non-MT applications (e.g., as a NL front-end to a virtual reality system), it is not the case that mathematical formalizations predict the specific combinations of spatial relations within natural language spatial predicates. Talmy (1983) captures this fact nicely in his detailed description of the meaning of the English word *across.*

2.5 Modularization of Knowledge Encodings

In order to talk about the encoding of spatial relations, we need to clarify which encodings appear in which part of the MT system. The following terms are used to classify the encoding of spatial relations on the basis of the "evidence" we have for them:

- *lexically explicit:* a spatial relation encoded explicitly in a word.[12]
- *lexically implicit:* a spatial relation encoded implicitly, or internal to the structure representing the meaning of a word.
- *logically inferable:* a spatial relation logically inferred from lexically explicit or implicit relations, but not itself part of the structure representing the meaning of a word.

In the first two cases, the relation appears in the lexical entry for the relevant word; in the third case, the relation does not appear in the lexical entry.

An example of the first case is the direction SOUTH as an abstract concept, which is lexically explicit in the word *south.*[13] An example of the second case is the direction UP as a lexically implicit component of the word *raise.* The implicit presence of this constituent is apparent in tests for synonymy: *they raised the platform one and a half feet*, *they raised the platform up one and a half feet.* Finally, as an example of the last category, the direction FROM and a location that is distinct from his home are logically inferable in the sentence *John arrived home*, where the verb's lexically implicit relation PATH contains the explicit

[11] We discuss our use of the terms *lexically implicit* and *lexically explicit* further in the next subsection.

[12] We use *word* here to refer to lexicon entry "names".

[13] The words in capital letters refer to the spatial relation, i.e., the abstract term.

DESTINATION *home*, and where we can infer logically that in a PATH ending at home, there was also a place from which the arriver, *John*, came.

The definition of these categories is tied to the way we have modularized PRINCITRAN into components. In the chart in Figure 4, the X's mark which types of encoding of a spatial relation may appear in which of the components in our MT system.

Spatial Relation	*Component of PRINCITRAN*		
	Syntax	IL	KR
lexically explicit	X	X	X
lexically implicit		X	X
logically inferable[14]			X

Figure 4: Encoding of Spatial Relations in the Components of PRINCITRAN

Following up on the examples above, the relation SOUTH in *south* will be represented at all levels in PRINCITRAN, whereas UP in *raise* will only be represented at the IL and KR levels, and FROM in *John arrived home* will only be represented at the KR level (as the result of inferencing).

We can readily see that the Syntax-IL mapping requires tracking which elements in the spatial predicates (at the IL level) appear in the surface SL and TL sentences and where in the sentence syntax they will be positioned. The KR-IL relation is not of this nature; rather, the KR system serves to verify the appropriateness of the spatial information that appears in the IL representation. One must not confuse the spatial information contained in the IL and that which is inferred by the KR system. To clarify this point, consider the following English sentences:

(5) (i) He took the book to Tanya's table
(ii) He took the book from Florence's floor

If the sentences are translated into German, the *take-to* component of the first sentence translates to *bringen* whereas the *take-from* component in the second sentence translates to *nehmen*. In both sentences there is an implicit PATH relation where a book moves from one location to another. The FROM direction is logically inferable in the first sentence but lexically explicit in the second sentence. The situation is reversed with a TO direction: the TO is lexically explicit in the first sentence, but only logically inferable in the second sentence. If our IL representation of the first sentence were to include the FROM relation by using a general PATH predicate with an inferred source location (here, not at Tanya's table)—and similarly if our IL representation of the second sentence were to include the TO relation by using a general PATH predicate with an inferred goal location (here, not at Florence's floor)—then at the point in translation where the system must generate a German sentence, it would have the added step of having to rederive what had been inferred (here,

[14] The logically inferable relations can be broken out into the "logically explicit" facts explicitly encoded in the KR system and the "logically implicit" facts that are derived from other facts and inference rules in the system.

the negated locations) and what had been in the SL lexicalization in order to select between the two German verbs.

This last example and the chart above help illustrate the double set of justifications that are required in a theory of the interlingua. In particular, the syntax-IL mapping provides one set of constraints on the IL, whereas the the IL-KR relation provides another set. Currently no theory of the interlingua defines these constraints and addresses the criteria to be used in evaluating them, yet it is clear from the above discussion that a non-uniform model does not hinder and indeed may actually facilitate capturing the types of distinctions (e.g. such as between sources and destinations in spatial paths) required for successfully translating diverse language pairs.

3 Defining an Interlingua

This section provides a specification for the interlingua. We start by outlining the resources that have contributed toward the development of an IL representation. We then define and illustrate the structure of the IL. Finally, we describe our use of the IL representation in interlingual MT, addressing issues of lexicon development and implementation status of PRINCITRAN.

3.1 Contributions to the LCS-derived Representation Language

Our IL representation derives from the lexical conceptual structures (LCSs) of Jackendoff (Jackendoff, 1983; Jackendoff, 1990; Jackendoff, 1991).[15] The LCS framework consists of three independent subsystems: fields, conceptual constituents, and boundedness and aggregation properties. Only the first two are currently a part of our IL framework.

The LCS *fields* (i.e., Loc(ational), Temp(oral), Poss(essional), Ident(ificational), Perc(eptual), and others) are motivated by well-known observations of *lexical parallelism,* where the same lexical item has parallel or related meanings in two or more semantic fields. The *conceptual constituents* in the second LCS subsystem are variants on predicate-argument structures. These include *primitive predicators* (such as GO, BE, and CAUSE) and their arguments and modifiers, each of which has an ontological *type* (e.g., Thing, State, Event, Place, Path, and Property); the internal structure of each constituent may decompose into another conceptual constituent. The primitive predicators are subscripted by *field* in addition to being typed by category.

As an example of how the primitive predicator GO (of ontological type Event) is used in representing the sentence semantics of a spatial expression (i.e., in the Loc(ational) field), consider the following case:

(6) (i) The ball rolled toward Beth.

(ii)
$$\begin{array}{l} [_{\text{Event}}\ \text{GO}_{\text{Loc}}\ ([_{\text{Thing}}\ \text{BALL}], \\ \qquad\qquad [_{\text{Path}}\ \text{TOWARD}_{\text{Loc}} \\ \qquad\qquad\qquad ([_{\text{Thing}}\ \text{BALL}], [_{\text{Position}}\ \text{AT}_{\text{Loc}}\ ([_{\text{Thing}}\ \text{BALL}], [_{\text{Thing}}\ \text{BETH}])])])] \end{array}$$

[15] See (Dorr, 1993) for details of a MT system whose IL is LCS-derived. Recently the LCS framework has been used by others for French, e.g., (Pugeault, Saint-Dizier, and Monteil, 1994; Verrière, 1994).

Roughly, this representation means "The ball went locationally toward Beth."[16] Predicators (enclosed in square brackets "[]") can take zero or more arguments (enclosed in parentheses "()"). In the lexical entry for the sense of *roll* in (6), the predicator GO takes two arguments: the first is a thematic (affected) object and the second is either a directional path or a means/manner by-phrase.[17]

Jackendoff makes the claim that the conceptual structures generalize across fields. In particular, he adapts a *localist* view, claiming that the formalism for encoding constituents in the spatial field at some level of abstraction, generalizes to other fields. A localist, or localist-related, approach to lexical semantics is by no means unique to Jackendoff. See, for example, among many others, Anderson (1971), Heine, Claudi, and Huennemeyer (1991), Langacker (1987), and Schank (1975). In particular, early translation approaches used the spatially oriented *conceptual dependency* (CD) representation as the basis for interlingual machine translation (Schank, 1975; Schank and Abelson, 1977; Lytinen and Schank, 1982). The CD-based IL is a decompositional representation based on a small set of primitives that revolve around basic spatial notions such as motion, location, and direction. However, the CD-based framework for MT does not provide a systematic relation between the interlingua and its corresponding surface realization. For example, there is no uniform mechanism for handling even simple translation divergences such as argument reversal cases (e.g., translation of the English sentence *I like Mary* into the Spanish sentence *me gusta María* ('Mary pleases me')). Our approach differs in that it provides a systematic syntax-IL interface geared toward providing a uniform treatment of translation divergences.[18]

Although Jackendoff addresses the problem of defining a mapping between the semantics (i.e., the IL) and its corresponding syntactic realization, his work does not address the **computational** issues associated with representing or processing LCSs.[19] In particular,

[16] Given that the ontological type is uniquely determined by the name of the primitive predicator, examples in later sections use a short-hand notation where the type is omitted.

[17] If the second argument is a means/manner by-phrase, it is still possible to instantiate a directional path in the IL representation, although this would not appear in the verb's lexical entry. We will clarify this point further shortly.

[18] If we dig a bit deeper into the reasons behind the difficulty of using the CD representation as an interlingua, we find that the fundamental problem is that the CD framework subscribes to the notion that every imaginable component of meaning must be captured in a single representational formalism. Indeed, the CD formalism is one of the most developed representations in the field of AI, with care given to including information beyond the scope of the LCS formalism, (e.g., concerning short-term and long-term memory). However, the CD approach pays a high price for incorporating "deeper" knowledge in a single representation without preserving structurally defined information. In addition to missing linguistic generalization in the IL-syntax mapping, the CD representation is difficult to bound: the decomposition process on which it is based may lead to deep recursion. As noted by Schank himself (1973, p. 201), this is particularly a problem with instrumentality:

> "If every ACT requires an instrumental case which itself contains an ACT, it should be obvious that we can never finish diagramming a given conceptualization. For [the] sentence [John ate the ice cream with a spoon], for example, we might have 'John ingested the ice cream by transing the ice cream on a spoon to his mouth, by transing the spoon to the ice cream, by grasping the spoon, by moving his hand to the spoon, by moving his hand muscles, by thinking about moving his hand muscles,' and so on ..."

[19] He points this out explicitly in response to criticism from some in the computational linguistic community

although Jackendoff writes that thematic relations (i.e., the roles in predicate-argument structure) depend crucially on an enriched KR, he leaves open to interpretation (i) what that KR ought to look like and (ii) what would constitute an adequate scheme for grounding the primitives of the LCS representation in the KR.

Another resource for the development of our IL representation is the linguistically motivated notion of *lexical semantic template* (LST) as defined in the work of Levin and Rappaport-Hovav (1995a; 1995b).[20] The LST framework provides a decomposition of verbs into *predicate* structure and non-predicate *constants*, where a verb with many meanings reflects the pairing of one constant with several different predicate structures. To illustrate this point, consider the following sentences:

(7) (i) The soldiers marched.
(ii) The soldiers marched to the barracks.
(iii) The soldiers marched clear of the falling rocks.
(iv) The soldiers marched the soles of their boots flat.
(v) The general marched the soldiers to the barracks.

In each of the sentences above, the verb *march* introduces a single constant, denoted with angle brackets, the manner-of-motion constant ⟨MARCH⟩. It is the predicate structure that distinguishes the verbs in each of these examples—respectively, simple manner of motion, directional motion, state-change resultative, "fake" reflexive resultative, and causative directional motion.[21]

This framework also allows the same predicate structure to be associated with different constant names; each predicate-constant combination is realized as a different verb in the surface form, but the argument-taking properties are identical:

(8) (i) They funneled the mixture into the jar.
(ii) They ladled the mixture into the jar.
(iii) They shoveled the mixture into the jar.
(iv) They spooned the mixture into the jar.

In all of the cases above, there is an affected object within a directional resultative, but the constant underlying each case, the instrument of the action, is different— respectively, ⟨FUNNEL⟩, ⟨LADLE⟩, ⟨SHOVEL⟩, and ⟨SPOON⟩. Note that such constants cannot be specified in isolation, but must be found in some KR.

It is precisely this LST-style systematic use of the constant in our representations that allows us to transcend Jackendoff's framework. For example, we have adapted his representation given previously in (6) to include the constant ⟨ROLL⟩, thus providing a means for grounding the representation in the KR:

(9) (i) The ball rolled toward Beth.

(Jackendoff, 1992).

[20] The name LST is taken from (Levin and Rappaport-Hovav, 1995a, p. 24).

[21] While *march* may appear in an infinite number of distinct semantic contexts, here we are only distinguishing among its structural frames for which there are syntactic diagnostics.

(ii) $$\begin{array}{l} [_{\text{Event}}\ \text{GO}_{\text{Loc}}\ ([_{\text{Thing}}\ \text{BALL}], \\ \qquad [_{\text{Path}}\ \text{TOWARD}_{\text{Loc}} \\ \qquad\qquad ([_{\text{Thing}}\ \text{BALL}], [_{\text{Position}}\ \text{AT}_{\text{Loc}}\ ([_{\text{Thing}}\ \text{BALL}], [_{\text{Thing}}\ \text{BETH}])])], \\ \qquad [\text{BY}\ \langle\text{ROLL}\rangle])] \end{array}$$

Roughly, this representation means "The ball went locationally toward Beth by rolling." Note that the primitive GO has two arguments in (6) but three arguments in (9). In the spirit of the LST framework, we adopt the former 2-argument representation for lexical items and the latter 3-argument representation for the IL form associated with the full sentence. In the lexical entry for *roll*, as used in the sentence above, the two arguments in the GO predicate are the thematic (affected) object and the means/manner by-phrase. The directional path predicate in the full sentential IL form above does not originate in the lexical representation for *roll*. It appears in the form above as the result of analyzing the phrase *toward Beth* within the verb phrase.

By adapting this representation, we retain the benefits of the IL-syntax interface, while addressing the computational issue of associating the IL primitives with concepts in the KR. Furthermore, we are able to preserve the structural constraints on the Loc(ational) field that hold in other fields. For example, in the Ident(ificational) field, sentence (10) will still have a structure parallel to the adapted one in (9) above.

(10) (i) The snowman melted into a puddle.

(ii) $$\begin{array}{l} [_{\text{Event}}\ \text{GO}_{\text{Ident}}\ ([_{\text{Thing}}\ \text{SNOWMAN}], \\ \qquad [_{\text{Path}}\ \text{TO}_{\text{Ident}} \\ \qquad\quad ([_{\text{Thing}}\ \text{PUDDLE}], \\ \qquad\qquad [_{\text{Position}}\ \text{AT}_{\text{Ident}}\ ([_{\text{Thing}}\ \text{SNOWMAN}], [_{\text{Thing}}\ \text{PUDDLE}])])], \\ \qquad [\text{BY}\ \langle\text{MELT}\rangle])] \end{array}$$

A third resource for our IL development is the semantic classification scheme of Levin (1993), which was developed as a foundation for determining shared syntactic behaviors across different verbs. In Levin's terms, a semantic class is a set of verbs where each member participates in the same set of surface alternations (i.e., syntactic behaviors). Levin suggests that members of a set share components of meaning; however, the predicate structures corresponding to meaning components are not identified. Levin's subsequent work with Rappaport-Hovav on the LSTs, already described above, also provides a more extensive linguistic analysis of verb behaviors. The semantic *categories* in this later work are more general than the semantic *classes* of Levin (1993). For example, the semantic category *verbs of motion* subsumes more than one of Levin's classes (*manner of motion*, *inherently directed motion*, etc.).

3.2 Decompositional/Predicate-Based Representation: RLCS

The combination of the three resources described above forms the basis for our lexical IL structure—called RLCS (*root word* LCS). Each RLCS is language-specific and is associated with one root form (e.g., uninflected form for verbs) in its language's MT lexicon. Consistent with the LST approach, verbs sharing a semantic class (from (Levin, 1993)) have the same predicate structure and differ only in the constant within that structure. All verbs within a given class are subject to the same set of linguistic constraints in syntax. Our IL structures diverge from Levin and Rappaport-Hovav's LSTs in terms of the particular primitive predicators and in our inclusion of semantic fields (Jackendoff, 1983). As our interest is in the area of spatial predicates, we have investigated verbs of *motion*, *removal*, and *placement*; however, the same linguistic constraints apply to verbs in other seemingly unrelated domains (such as *sound*).

As outlined informally above, we distinguish among the three types of primitives within our RLCSs:

- Constants: within-class identifiers (e.g., ⟨ROLL⟩) which are distinct from predicate structure, but are attached within an IL form.
- Situation-Level Predicators: a small set of non-constant primitives corresponding to primary relations (e.g., GO, ACT, BE, CAUSE).[22] These primitives are associated with a small set of fields denoted by subscripted labels (e.g., Loc(ational), Temp(oral), Poss(essional), Ident(ificational), Perc(eptual)).
- Situation-Internal Predicators: primitives (e.g., TOWARD, AT) dominated by situation-level predicators.

Both Situation-Level and Situation-Internal predicators (enclosed in square brackets "[]") can take zero or more arguments (enclosed in parentheses "()"). With respect to constants, each word sense uses a constant to encode information that is idiosyncratic to (but not necessarily unique to) that particular word. The constant is also grounded as a concept of the KR system.

We note that the constant in a word's RLCS is serves as a link between that word's lexical-semantic structure, which maps systematically to syntax, and conceptual knowledge that is required for logical inferencing. This is consistent with the work of Levin and Rappaport-Hovav (1995b) and Grimshaw (1994), where *semantic structure*—the structural component of meaning which maps systematically to syntax—is distinguished from *semantic content*—the idiosyncratic component of meaning which has its underpinnings in conceptual knowledge. These terms readily extend to our framework: the IL predicates are a part of the *semantic structure,* serving as a link between the IL and the syntax, and the constants such as ⟨ROLL⟩ are a part of the *semantic content,* conveying idiosyncratic knowledge necessary for distinguishing this word from others within the same semantic class. At generation time when the MT system accesses the *IL lexicon*—a reverse index into TL lexicons (as discussed in section 5)—the constants are critical to handling translation mismatches such as lexical gaps in the target language.[23]

The systematic use of constants here is an improvement on earlier work (see, e.g., (Dorr, 1993)) where manner constants such as *runningly* were viewed on a par with primitives such

[22] The primitive ACT has been borrowed from the LST framework in order to characterize certain *activities* (such as *shovel*) which could not be adequately characterized by Jackendoff's GO primitive. In his more recent LCS specification, Jackendoff (1990) augments the *thematic* tier of his earlier work (Jackendoff, 1983) with an *action* tier which serves to characterize activities using additional machinery. We choose to simplify this characterization by using the ACT primitive rather than introducing yet another level of representation.

[23] Note that the constant is not always identified by the word itself, but by some form that is semantically related to the verb.

as GO and CAUSE; these "constants" were proliferated, without linguistic justification, throughout the lexical IL representations. Typically, such usage was accompanied by a footnote indicating that further investigation into the nature of lexical-semantic structure would be necessary. We attempt to remedy this shortcoming by developing our current representations in terms of lexical IL templates so that they conform to structural constraints invoked by the LST framework.

Within this "lexical IL template" approach, lexical entries adhere to a small set of structural requirements that divide verbs into "result verbs" such as *clear*, which incorporate a resultant state, and "manner/means verbs" such as *shovel*, which incorporate manner or means. The "result" vs. "means/manner" dichotomy is a basic distinction in the verb lexicon that has been carefully studied and characterized by Levin and Rappaport-Hovav (1995a; 1995b). The distinction cuts across verbs in the spatial domain in our investigation as well as across verbs from seemingly unrelated semantic classes.[24]

The reason the "result" vs. "means/manner" distinction is important is that it correlates with a difference in syntactic behavior. As an example, consider the *removal* category. Within this category, the behavior of verbs in one class, such as *clear,* contrasts syntactically with verbs in another class, such as *shovel*. Only in the case of *shovel* is the change-of-state resultative construction allowed:[25]

(11) (i) ∗I cleared the table clean

(ii) I shoveled the driveway clean

On the other hand, an "of" phrase can be used for *clear*, but not *shovel*:

(12) (i) I cleared the table of dishes

(ii) ∗I shoveled the driveway of snow

Figure 5 shows five broad semantic categories (taken from Levin and Rappaport-Hovav (1995b)) and, for each, two example verbs along with their associated RLCS representations in the lexicon. Specifically, each broad category includes one "result" verb entry (e.g., *leave*) and one "means/manner" verb entry (e.g., *run*). Class numbers from Levin's 1993 book are provided for each verb example. The distinction between these two types of verbs is reflected in the constant of the RLCS representation. For result verbs, the constant corresponds to a resulting state (as in *clear* and *fill*) or location (as in *leave*). For manner/means verbs, the constant may specify a manner (as in *run* and *pour*) or means (as in *shovel*).

[24] To simplify the discussion here, we have restricted our examples to verb senses that fall into one of the two categories, either result or means/manner. There are also some verbs that have both elements of meaning, such as *cut*.

[25] Occasionally it is possible for a result verb to participate in a resultative construction, but only if the result phrase further specifies the verb's inherent result as in *I drained the tub dry*. In such "non-canonical" examples, we would expect human judgments to vary with respect to grammaticality.

[26] The *HEAD* symbol is a place-holder that points to the root, event-level node of the overall lexical entry. Modifiers, such as instrumental phrases, typically include this symbol. See (Dorr, 1993) for more details of this notation.

Category	**Verb**	**Class**	**RLCS**	**Template**
Motion	leave	51.2	[GO$_{\text{Loc}}$ (Y, [⟨AWAY-FROM⟩$_{\text{Loc}}$ (Y, [AT$_{\text{Loc}}$ (Y, Z)])])]	(13)(i)(b)
	run	51.3.1	[ACT$_{\text{Loc}}$ (X, [BY ⟨RUN⟩])]	(13)(ii)(a)
Removal	clear	10.3	[CAUSE (X, [GO$_{\text{Ident}}$ (Y, [TOWARD$_{\text{Ident}}$ (Y, [AT$_{\text{Ident}}$ (Y, [⟨CLEAR⟩$_{\text{Ident}}$ ([⟨OF⟩$_{\text{Poss}}$ (*HEAD*, Z)])])])])])])][26]	(13)(i)(a)
	shovel	10.4.2	[ACT$_{\text{Loc}}$ [ON$_{\text{Loc}}$ (Y)], [BY ⟨SHOVEL⟩])]	(13)(ii)(a)
Placement	fill	9.8	[CAUSE (X, [GO$_{\text{Ident}}$ (Y, [TOWARD$_{\text{Ident}}$ (Y, [AT$_{\text{Ident}}$ (Y, [⟨FULL⟩$_{\text{Ident}}$ ([⟨WITH⟩$_{\text{Poss}}$ (*HEAD*, Z)])])])])])])]	(13)(i)(a)
	pour	9.5	[ACT$_{\text{Loc}}$ [ON$_{\text{Loc}}$ (Y)], [BY ⟨POUR⟩])]	(13)(ii)(a)
Sound	say	37.7	[CAUSE (X, [GO$_{\text{Ident}}$ (Y, [TOWARD$_{\text{Ident}}$ (Y, [AT$_{\text{Ident}}$ (Y, [⟨SAID⟩$_{\text{Ident}}$])])])])])]	(13)(i)(a)
	shout	37.3	[ACT$_{\text{Perc}}$ [ON$_{\text{Perc}}$ (Y)], [BY ⟨SHOUT⟩])]	(13)(ii)(a)
Killing	kill	42.1	[CAUSE (X, [GO$_{\text{Ident}}$ (Y, [TOWARD$_{\text{Ident}}$ (Y, [AT$_{\text{Ident}}$ (Y, [⟨KILLED⟩$_{\text{Ident}}$ ([⟨WITH⟩$_{\text{Instr}}$ (*HEAD*, Z)])])])])])])]	(13)(i)(a)
	stab	42.2	[ACT$_{\text{Perc}}$ [ON$_{\text{Perc}}$ (Y)], [BY ⟨STAB⟩])]	(13)(ii)(a)

Figure 5: RLCSs Based on Levin's Verb Classification

We use the following basic templates, cross-referenced in figure 5, to characterize the result vs. manner/means dichotomy:[27,28]

(13) (i) **Result Verbs:**

(a) $\underline{\text{[CAUSE (X,}}\ \text{[GO}_{\text{Ident}}\text{ (Y, [TOWARD}_{\text{Ident}}\text{ (Y, [AT}_{\text{Ident}}\text{ (Y, [}\langle\text{STATE}\rangle\text{])])])}\underline{\text{)]}}$

(b) $\underline{\text{[CAUSE (X,}}\ \text{[GO}_{\text{Loc}}\text{ (Y, [}\langle\text{DIRECTION}\rangle_{\text{Loc}}\text{ (Y, [AT}_{\text{Loc}}\text{ (Y, Z)])])}\underline{\text{)]}}$

(c) $\underline{\text{[CAUSE (X,}}\ \text{[GO}_{\text{Loc}}\text{ (Y, [TOWARD}_{\text{Loc}}\text{ (Y, [}\langle\text{POSITION}\rangle_{\text{Loc}}\text{])])}\underline{\text{)]}}$

(ii) **Means/Manner Verbs:**

(a) $\text{[ACT}_{\text{Loc/Perc}}\text{ (X, }\underline{\text{[ON}_{\text{Loc/Perc}}\text{ (Y)],}}\text{ [BY }\langle\text{MEANS/MANNER}\rangle\text{])]}$

(b) $\underline{\text{[CAUSE (X,}}\ \text{[GO}_{\text{Loc/Perc}}\text{ (Y, }\underline{\text{[TOWARD}_{\text{Loc}}\text{ (Y, [AT}_{\text{Loc}}\text{ (Y, Z)])],}}$
$\text{[BY }\langle\text{MEANS/MANNER}\rangle\text{])}\underline{\text{)]}}$

The underlining in (13) is a shorthand notation to condense the presentation of multiple templates. Note that these basic templates vary along a number of different dimensions, e.g., the name of the situation-level predicator, the number of arguments associated with the predicator, the allowable fields, and the position of the constant. In addition, certain primitives are interchangeable with others, e.g., AT_{Loc} can be replaced with other static spatial relations, such as the static UP_{Loc} (meaning "at the top of") in the RLCS of the transitive verb *climb* (as we will see in section 4).

In short, these basic templates show how the full space of RLCSs vary predictably and capture the "result" vs. "means/manner" distinction, thus enhancing our approach to defining a systematic interface between the IL and the syntactic structure. This interface is the topic of the section 4.

3.3 Use of the RLCS in Interlingual MT

The representations and constraints described above serve as the foundation of a large-scale, RLCS-based lexicon for interlingual MT of English, Arabic, French, Korean, and Spanish. We have built a database of English RLCS representations for the 3828 verb entries in Levin (1993), where an *entry* is a "semantic class/verb sense" pair, e.g., "42.1/kill" from Figure 5. These RLCSs were automatically derived from a manually-encoded set of templates for all 192 of Levin's semantic classes.[28] The classes consist of a total of 2775 unique verbs distributed across 3828 Levin-based entries (some verbs occur in multiple classes). In addition, we have developed RLCS representations for 3500 non-Levin verbs,[29] by associating with each verb a semantic class and then instantiating the RLCS template associated with that class in order to build that verb's lexical entry. For details of experiments in the automatic construction of an RLCS-based lexicon for other languages, e.g., Arabic, see Dorr, Garman, and Weinberg (1995).

[27] For the purpose of this discussion, we restrict our attention in these templates to non-stative verbs. States, however, are also allowed through the use of the BE situation-level predicator.

[27] Templates (13)(i)(c) and (13)(ii)(b) are not included in figure 5; examples of these will be given in section 4.1.

[28] Levin's classes are labeled with numbers ranging from 9 to 57. However, the actual number of semantic classes is 192 (not 46) due to many class subdivisions under each major class.

[29] By non-Levin verbs, we mean verbs or individual verb senses that do not appear in Levin (1993).

The full set of RLCS's has now been ported to Arabic and Spanish; these are being used in a large-scale "analysis component" of an implemented foreign language tutoring system (Dorr et al., 1995). The "analysis component" of this tutoring system, coupled with the RLCS lexicon, serve as the core components of PRINCITRAN. A prototype version of the generation component has been constructed for complete end-to-end translation (Dorr et al., 1994).

Our interlingual model assumes that the analysis phase (from SL input to IL form) involves the SL lexicon and the generation phase (from IL form to TL output) involves the TL lexicon. The role of the RLCS's in each of the lexicons is slightly different, as the task in each direction is inherently distinct: (i) during analysis, the appropriate RLCSs are selected from the SL lexicon, as a function of syntax, and the IL is constructed compositionally from the predicate/constant information in the RLCSs; (ii) during generation, the RLCS's of the TL are accessed from a larger *IL lexicon* (described in section 5) and the TL words are selected such that their RLCSs semantically cover the predicate/constant information in the IL. The next section examines each of these two tasks more closely, at the interface level, where the RLCS serves as the primary mediating structure.

4 The Syntax/IL Interface

As illustrated in the previous section, our investigation into the nature of spatial mismatches has led to a better understanding of what must necessarily be included in the IL. This section describes the interface between the IL representation and the syntax, and in particular, discusses the allocation of information in the RLCS-based lexicons and the rules that operate on RLCSs to produce the full (sentential) IL.

4.1 Information Allocation in RLCS-Based Lexicon

In Section 2.4 we argued that, in order to provide adequate cross-linguistic coverage, a MT system must address cases where languages differ with respect to their patterns of lexicalization, e.g., which arguments may be incorporated into a lexical item (Talmy, 1983; Talmy, 1985). An IL formalism that identifies incorporated information will provide a greater variety of lexicalization options for mapping between the source and target language. The decompositional nature of our RLCS representation provides a means for capturing incorporated information.

Consider, for example, the English verbs *lift*, *ascend*, and *climb* as used in the following sentences:

(14) (i) Mary lifted the baby.

(ii) The plane ascended.

(iii) John climbed the stairs.

All three verbs hide (or incorporate) the semantics of another word, *up*. This incorporated information is encoded in each verb's lexicon entry (i.e., in the RLCS) either as its own constant or as part of the meaning of a conceptually more complex constant:[30]

[30] These are cross-referenced with the corresponding templates from (13).

(15) (i) lift: $[\mathrm{CAUSE}\ (\mathrm{W}, [\mathrm{GO}_{\mathrm{Loc}}\ (\mathrm{Y}, [\mathrm{BY}\ \langle \mathrm{LIFT} \rangle])])]$ (13)(ii)(a)

(ii) ascend: $[\mathrm{GO}_{\mathrm{Loc}}\ (\mathrm{Y}, [\mathrm{TOWARD}_{\mathrm{Loc}}\ (\mathrm{Y}, [\langle \mathrm{UP} \rangle_{\mathrm{Loc}}])])]$ (13)(i)(c)

(iii) climb: $[\mathrm{GO}_{\mathrm{Loc}}\ (\mathrm{Y}, [\mathrm{TOWARD}_{\mathrm{Loc}}\ (\mathrm{Y}, [\mathrm{UP}_{\mathrm{Loc}}\ (\mathrm{Y}, \mathrm{Z})])], [\mathrm{BY}\ \langle \mathrm{CLIMB} \rangle])]$ (13)(ii)(b)

During sentence analysis, the constants in *lift, ascend,* and *climb* encode our notion of inherent upward motion in distinct ways.[31] In the *lift* RLCS, the constant encodes that motion. It is the ⟨LIFT⟩ constant that allows the *lift* RLCS to be composed with the particle *up*. That the semantics of upward motion belongs in a constant that in turn is grounded in the KR may seem too well hidden in the representation. We note here only that, for many native speakers of English, there is no strict requirement that a lifting motion necessarily be upwards; for example, one person could be on a ladder removing items from the top shelf of a kitchen cabinet and be *lifting down* those items to someone standing on the ground beneath them.

The *ascend* RLCS differs from the one for *lift* in two ways. First, *lift* in the sentence above is a causative verb, requiring an external cause; it contains the primitive CAUSE and an argument W. The verb *ascend* is not grammatically causative: in English, we cannot say that *X ascends Y* means *X caused Y to go upward.* Second, unlike *lift*, *ascend* may not co-occur with a spatial direction such as *down*: **he ascended down the hill.* Thus, we say *ascend* is inherently directed with TOWARD and the constant ⟨UP⟩.

The RLCS for *climb,* differs from that of *lift* and *ascend* in that there is an UP predicator with two arguments, the second of which must be lexicalized. The particular meaning of *climb* that comes with an obligatory argument as theme incorporates UP as its own primitive. That is, *climbing* any object entails upward motion on that object. The UP is disjoint from the ⟨CLIMB⟩ constant which is reserved for the clambering manner of motion that defines *climbing* in English. It is interesting to note that in languages as distinct as German and Turkish, one cannot use the same "climb" word for expressing *climbing a mountain* (inherently upward) and *climbing down a mountain.* That is, where we have two RLCSs for two distinct uses of the one English word *climb*, other languages have two separate words.[32,33]

[31] For a more detailed examination of the computational lexical semantics of "up" in English and Turkish, see Voss, Dorr, and Şencan (1995).

[32] If the target language were German, for either a *lift* or *lift up* SL input, there would be a *heben/anheben* choice (where the German prefix *an* corresponds loosely to the English *up* in this case). For either a *climb* with direct object in the SL input, there would be a *steigen/aufsteigen* choice (among a few others). And finally for an *ascend* SL input, there would be a *steigen/ersteigen* choice (among possibly a few others). These choices correspond to the lexical synonymy case in Figure 3.

[32] A reviewer points out that one needs to allow for different verb-particle realizations:

(i) John climbed/ascended the rope.

(ii) John climbed/*ascended up the rope.

The mechanism that achieves this distinction lexically is the star-marker, described in (Dorr, 1993). The star-marked constituents would be ⟨LIFT⟩ in (15)(i) and $\mathrm{UP}_{\mathrm{Loc}}$ in (15)(iii). By contrast, the $\mathrm{UP}_{\mathrm{Loc}}$ would not be star-marked in (15)(ii).

4.2 Composition of the IL

As we noted above, the constants and primitives in the IL forms corresponding to *lift, ascend,* and *climb* are all retained into their composed IL forms. This illustrates a key benefit of our formalism: the substructure and constant information of the SL input is preserved and so the decision process on lexicalization into the TL is not prematurely foreclosed. By leaving the lexicalization decision open into the TL phase, our MT system allows for TL-specific pragmatic information to be used and for stylistic choices to be made in the final generation steps[33]—after the TL lexical options have been identified from the IL form.

We are now prepared to describe the composition process that derives the IL representations. During the analysis phase of translation, the RLCSs are composed, producing what is called a CLCS (for *composed* LCS) as the IL form. This process is governed by linguistic rules that are associated with the RLCSs of the SL sentence. We assume that verbs, or more accurately verb senses, sharing a semantic class have the same RLCS template and that all RLCSs in a given class are subject to the same set of constraints on possible compositional operations. An example of a compositional operation is one that derives the change-of-state resultative construction illustrated in (11) above; this operation applies to verbs like *shovel* (in class 10.4.2 of Levin (1993)) but not verbs like *clear* (in class 10.3 of Levin (1993)).

Recall that our motivation for adhering to certain restrictions of the LST framework was that it provides a means for defining a systematic interface between IL forms and their possible syntactic realizations. In particular, we exploit the constraints inherent in the LST framework for distinguishing between different syntactic behaviors as exemplified above in (11) and (12). There are two advantages to using the LST framework for defining the RLCS lexicon: (i) Templates may be freely augmented by composition with other templates, subject to certain constraints; (ii) Constants in the templates are systematically related to surface lexical items and arguments in the templates are systematically related to syntactic argument positions.

To illustrate these two points, consider the different syntactic distributions of *clear* and *shovel* again. The lexical entry for *clear* directly incorporates the resultant state ⟨CLEAR⟩ whereas no such state is available in the *shovel* entry. (These entries are shown in figure 5.) Thus, there is no way to further augment the *clear* template through a change-of-state resultative composition operation as in the sentence **I cleared the table clean*. In the case of *shovel*, this option is open: *I shoveled the driveway clean*. The operation that derives this sentence yields the following CLCS representation:

(16) I shoveled the driveway clean.
[CAUSE
([ACT$_{\text{Loc}}$ (I, [ON$_{\text{Loc}}$ (DRIVEWAY)], [BY ⟨SHOVEL⟩])],
[GO$_{\text{Ident}}$
(DRIVEWAY,
[TOWARD$_{\text{Ident}}$ (DRIVEWAY,
[AT$_{\text{Ident}}$ (DRIVEWAY, [⟨CLEAN⟩$_{\text{Ident}}$])])])])]

Optionally, we could add an "of" phrase, as in the sentence *I shoveled the driveway clean of snow*. The word "of" corresponds to constant, ⟨OF⟩, which is available as an inherent

[33] For example, the TL discourse may be informal and call for a *go up* in lieu of an *ascend*.

argument of ⟨CLEAN⟩:

(17) I shoveled the driveway clean (of snow).
[CAUSE
([ACT$_{\text{Loc}}$ (I, [ON$_{\text{Loc}}$ (DRIVEWAY)], [BY ⟨SHOVEL⟩])],
[GO$_{\text{Ident}}$
(DRIVEWAY,
[TOWARD$_{\text{Ident}}$ (DRIVEWAY,
[AT$_{\text{Ident}}$ (DRIVEWAY, [⟨CLEAN⟩$_{\text{Ident}}$ ([⟨OF⟩$_{\text{Poss}}$ (*HEAD*, SNOW)])])])])])]

Implicit in the machinery described here is the assumption that all overt lexical items in the surface sentence must necessarily be related to a constant, whether inherent or compositionally introduced. This assumption rules out certain syntactic realizations for *shovel* that are available for verbs like *clear*. For example, *clear*, allows a bare "of" phrase to be used, as in *I cleared the table of dishes*, since this option is available in the lexical entry. This option would produce the following representation:

(18) I cleared the table of dishes.
[CAUSE
(I,
[GO$_{\text{Ident}}$
(TABLE,
[TOWARD$_{\text{Ident}}$ (TABLE,
[AT$_{\text{Ident}}$ (TABLE, [⟨CLEAR⟩$_{\text{Ident}}$ ([⟨OF⟩$_{\text{Poss}}$ (*HEAD*, DISHES)])])])])])]

The words *clear* and *of* in the surface sentence are mapped into the constants ⟨CLEAR⟩ and ⟨OF⟩ in this representation.

The same is not true of *shovel*, i.e., it would not be possible to introduce a bare "of" phrase, as in *I shoveled the driveway of snow*:

(19) I shoveled the driveway clean (of snow).
[CAUSE
([ACT$_{\text{Loc}}$ (I, [ON$_{\text{Loc}}$ (DRIVEWAY)], [BY ⟨SHOVEL⟩])],
[GO$_{\text{Ident}}$
(DRIVEWAY,
[TOWARD$_{\text{Ident}}$ (DRIVEWAY,
[AT$_{\text{Ident}}$ (DRIVEWAY, [⟨???⟩$_{\text{Ident}}$ ([⟨OF⟩$_{\text{Poss}}$ (*HEAD*, SNOW)])])])])])]

Here, the structure is ruled out because there is no word in the surface sentence to map onto a state constant in the underlying representation. Note that this constraint is inherent in the machinery of the LST framework; it falls out from restrictions on the representation itself, not from the application of explicit constraints. This is precisely the benefit we seek to gain in adapting this framework to the the RLCS framework described above.

In short, the lexical representation of a verb in the CLCS, i.e., after composition, does *not* differ from that of its original form in the RLCS, i.e., before composition. In other words, the substructures within a verb are retained, thus (monotonically) preserving the predicates for the later TL generation phase.

4.3 Cross-Linguistic Application of the RLCS Framework

We turn now to the use of this representational approach in an interlingual MT system. It is clear that there would be no benefit to basing our RLCS representations on the LST framework unless there were some way of accounting for cross-linguistic variation with respect to the applicability of composition operations. For example the change-of-state resultative construction, as in *I shoveled the driveway clean*, is not available in a language like Spanish; thus, the same meaning must be conveyed either as a paraphrase with two clauses or as a single clause with some omitted or redistributed information:

(20) (i) Traspalé el garaje para que esté limpio
'I shoveled the driveway so that it is clean'

(ii) Limpié el garaje (con una pala)
'I cleaned the driveway (with a shovel)'

The approach we adopt to handling such a distinction is to parameterize the IL/syntax mapping so that the change-of-state resultative operation is applied in English, but not in Spanish. For each natural language, we identify allowable sets of composition operations in that language, and then we identify, more narrowly, the classes of lexical items to which these operations apply. The parameterization permits each language to have its own idiosyncratic syntactic realization while sharing the same underlying IL representations.

The first sentence above most closely reflects the meaning of the English change-of-state resultative construction; thus its CLCS will be identical to that of the English sentence. The second sentence, on the other hand, will map to a different CLCS. The two respective CLCS representations are given here:

(21) (i) [CAUSE
([ACT$_{\text{Loc}}$ (I, [ON$_{\text{Loc}}$ (DRIVEWAY)], [BY ⟨SHOVEL⟩])],
[GO$_{\text{Ident}}$
(DRIVEWAY,
[TOWARD$_{\text{Ident}}$ (DRIVEWAY, [AT$_{\text{Ident}}$ (DRIVEWAY, [⟨CLEAN⟩$_{\text{Ident}}$])])])])]

(ii) [CAUSE
(I,
[GO$_{\text{Ident}}$
(DRIVEWAY,
[TOWARD$_{\text{Ident}}$ (DRIVEWAY, [AT$_{\text{Ident}}$ (DRIVEWAY, [⟨CLEAN⟩$_{\text{Ident}}$])])])],
[⟨WITH⟩$_{\text{Instr}}$ (*HEAD*, ⟨SHOVEL⟩)])]

Interestingly, the difference between the two translations is one of style; in Spanish, the first is much more awkward than the second.[34] Yet it is the first one that fits our translation scheme, in going from English to Spanish. Given that the parameterized scheme disallows the change-of-state resultative construction in Spanish, the syntactic realization of the first structure would result in the two-clause realization, *Traspalé el garaje para que esté limpio*.

[34] We have been informed by native speakers that climate differences make shoveling a driveway a rarety in many Spanish-speaking countries. Thus, the second sentence is also somewhat awkward, but for different reasons; the same sentence with the word *broom* (instead of *shovel*) is much more natural sounding.

The second CLCS, on the other hand, would never be the result of analysis of the original English sentence, *I shoveled the driveway clean*. The question of whether we could view this alternate structure—which actually contains the main sub-components of the first structure—as an adequate match for translation into Spanish depends heavily on whether it is possible to omit or redistribute different pieces of information based on the context of the sentence. For example, if one can determine from the setting of the sentence that the activity causing the clean state of the driveway involves a shovel, then perhaps this piece can either be dropped or moved into an instrumental position. Such an issue has been discussed at length by Slobin (forthcoming), where English and Spanish are distinguished by their inclusion of context information. In fact, according to Slobin, Spanish translators often make changes that *reduce* the amount of information conveyed in the English sentence in order to produce the Spanish sentence. Currently, our system opts for the more awkward translation—as long as it is syntactically realizable—over a loss or redistribution of information in the final output. However, it is clear that, with a richer theory of context, we have enough machinery to generate a TL sentence that corresponds to the second structure rather than the first one.

Note, by contrast, that a translation in the other direction (from Spanish to English) would most likely involve the second structure as the source-language sentence since a native speaker typically would not start off with the more awkward choice. In this case, perhaps it would be reasonable to redistribute or even *augment* the information (if the instrument is omitted) in the interlingua in order to produce an English output. According to Slobin, English translators sometimes do add a bit of information, but in most cases, they follow the original Spanish sentence in producing the English sentence. This is fortunate for us since, as can be seen in the structures in (21), it would be a simple task to map the first to the second, assuming the instrument can be dropped (since this would require reduction of information), but not the other way around (since this would require addition or redistribution of information). On the other hand, if we were to couple a rich theory of context with our rules of composition (which, for example, would allow change-of-state resultatives to be expressed in English), then we could perhaps identify cases where a more elaborate English sentence could be produced from a simpler CLCS form.

The RLCS scheme described here has the added benefit that these very same structures are well-interfaced with the KR module as described in the next section.

5 The IL/KR Interface

The previous section described the relation between the IL and the syntax, showing that the MT task is not hindered by taking a non-uniform approach to defining the representational languages at these two levels. This section further demonstrates this point by describing a systematic relation between the IL and the KR: the IL primitives in the IL forms are grounded as KR concepts, making deeper conceptual knowledge stored in the KR accessible in processing the IL forms. As above, we show that the division into levels is both linguistically motivated and applicable to diverse languages.

Here we present the IL/KR interface in terms of three MT system processing steps where the IL and KR "meet" at runtime: during IL filtering, IL composition in the analysis phase, and IL decomposition in the lexical selection of the generation phase. Our goal is to convey broadly the distinct range of issues that arise in defining the "meeting" at the interface in each of these steps. In the first subsection, we consider *object-specific* knowledge in the KR and its role in the filtering phase, eliminating IL forms that were composed during the analysis phase but that are incompatible with known spatial properties of objects. In the following subsection, we examine the role *event* knowledge in the KR can play when IL composition fails during the analysis of the SL input. In the final subsection, we focus on the generation phase and the role of the KR system in holding the *IL lexicon,* a reverse index for retrieving lexical entries from TL lexicons.

For the examples in the first two subsections, the "meeting" of the IL and KR arises when the MT system examines an IL primitive in an IL form (eg., in comparing two IL forms for type matching during IL composition) and then looks into the KR ontology for its grounded version of that primitive. In these cases, the leap, as it were, by the MT system from the IL primitive to the KR primitive may be done solely by looking up that term in the KR ontology. There is no structured connection or link between the IL primitive in the IL form composed at runtime and the KR primitive with further knowledge about that IL primitive. By contrast, in the third subsection, the "meeting" of the IL and the KR primitives takes place within the KR system itself: the IL lexicon, a structure accessing all lexical IL forms, or RLCSs, is constructed as its own hierarchy within the KR system, enabling the primitives within those entries to be defined in terms of the KR primitives in the KR system's separate KR ontology.

Although we have placed the IL lexicon within the KR system, as will be described below, we need to reiterate here that the IL and KR are, nonetheless, two distinct levels of representation. The syntax of the IL forms differs from that of expressions in the KR system and the set of primitives in the IL forms is a proper subset of the concepts in the KR. We take the IL lexicon to be the lexicalization space of the MT system, delimiting the set of possible lexical IL forms. By contrast, we take the KR ontology—a data structure in the KR component that is distinct from the IL lexicon—to be the concept space of the MT system, establishing the set of ontological terms in which the IL lexicon's primitives may be grounded. While all primitives in lexical IL forms correspond to KR concepts, not all KR concepts have a corresponding entry in the IL lexicon.

We should note that, not being the developers of the KR systems we have worked with, but rather KR consumers, we found that, as our MT system evolved, so did our KR needs. Our initial choice for a KR system, PARKA, was adequate for the problems we tackled in the filtering phase and in the analysis phase during composition failures.[35] For the generation phase, however, we moved to LOOM with its more developed user interface, documentation, and its classifier, an option not available in the earlier system. At this point our work has taken the form of developing experimental KR ontologies in PARKA and LOOM, and most recently, the testing of the IL lexicon within LOOM.

[35] PARKA is a research KR system under development at the University of Maryland.

5.1 During IL Filtering Phase

We make the assumption that object-specific knowledge, i.e., non-linguistic information about physical objects in the real world, belongs at the conceptual level in the KR component of the MT system. Given this *division of labor,* here we will look at an example where object knowledge is used after the full IL form for the SL input has been composed.

First consider the translation of the following English sentence into German:

(22) English: The mouse ran under the table

(i) German: Die Maus ist unter dem Tisch gelaufen
'The mouse ran (about in the area) under the table'

(ii) German: Die Maus ist unter den Tisch gelaufen
'The mouse ran (to a place somewhere) under the table'

(iii) German: Die Maus ist unter dem Tisch durch gelaufen
'The mouse ran (past a place somewhere) under the table'

During sentence analysis, two syntactic structures are created, capturing the fact that the phrase *under the table,* can attach either as an adjunct (in case (22)(i)) or as an argument (in cases (22)(ii) and (22)(iii) to the verb). At the semantic interpretation step, the analysis creates another split in the interpretation, differentiating between two senses of the PP argument as a PATH (cases (22)(ii) and (22)(iii)). From the three IL forms created in the analysis phase, the MT system then in the generation stage, outputs a distinct German sentence.

Now consider the translation of a slight variant on the English sentence:

(23) The mouse ran under the fence

The change from *table* to *fence* in sentence (23) does not alter the processing steps in the MT system's analysis phase from those described above for sentence (22). However, since the difference between tables and fences is only captured at the KR level in our system, it is consistent with our division of labor that the MT system should use this knowledge during its filtering phase when KR-based information is synthesized up through the IL form. In particular, the system must filter out the interpretations, paralleling cases (22)(i) and (22)(ii) above, of the mouse running about in the area under the fence and the mouse running to a place under the fence.

The key to solving this problem is developing representations for tables and fences that are adequate for differentiating their prototypical spatial properties when it comes to motion of an object on a horizontal plane under them. Our first-pass approach to this problem was to use the feature-based notation developed by Jackendoff (1991). This choice was most attractive because the features were defined within his LCS framework and his examples provided clear starting points for our own test sentences. We were able to apply the feature system to capturing prototypical boundedness properties of physical objects, including tables and fences, and path types and places. Then, by extending the features as well as to physical motions, the features in the different KR concepts were available across the relevant IL types (THING, PATH, PLACE, and EVENT). Rules for synthesizing the features up through the

IL forms were developed so that conflicting values signaled an invalid IL form that was not acceptable for translation. For example, the bounded 2-dimensional spatial relation of *under* type PLACE conflicted with the unbounded 1-dimensional schematization of a fence, in synthesizing the KR feature values grounding IL primitives in the IL predicate for *under the fence* in the PP adjunct analysis. With these adjustments, the filtering phase eliminated out of two of the three interpretations in the sentence (23) as needed.

This test effort also pointed out several limitations to the feature system in our task. We found it difficult to agree among ourselves on the features for an object independent of any spatial relation. For example, when coding the features we needed to designate one of the several schematizations of the object as primary. When we knew which spatial relation was going to be evaluated on that object, as was the case in our test work, the selection of the primary schema was straightforward. Without that information, our approach led only to an arbitrary ordering of one schema over another, an unsatisfactory engineering decision at best. Of equal concern was the restricted notation available in Jackendoff's approach to physical motion. In particular, while the features—developed for capturing linguistic *aspect*—differentiated bounded paths from unbounded ones, as in (22)(ii) versus (22)(iii) above, they were not expressive enough to contrast the wider range of spatial motions, as found in the manner of motion and directed verb classes of Levin (1993).

Despite the limitations of this specific object representation system, the test experiment did show that filtering in our MT system based on KR-based object information was possible. With further work now underway in integrating natural language and vision research, the field of AI may provide us with a more adequate formalism in the near future for further testing our system.

5.2 During IL Composition

When IL composition for a syntactic parse succeeds, the analysis phase is complete for that parse and the filtering phase begins. However when the IL composition fails in certain well-defined ways on a particular IL form under construction, the MT system can recover the semantic interpretation by accessing information stored in KR concepts. This section looks at one such situation, when a type mismatch occurs between an adjunct and the site at which it attaches.

We make the assumption in the lexical IL forms, or RLCSs, for verbs that the type of the *argument* position(s) is coded on a verb-by-verb basis, but that the *adjunct* position is always of type PLACE. That is, in the interpretation of our IL forms, the adjunct position corresponds to the spatial location, or frame location, of the event represented by the verb. For example, in the sentence (24), the PP *on the bed* may be attached as the verb's argument in which case the sentence refers to the specific manner of motion meaning in case (24)(i). When the PP is attached as the verb's adjunct, it may only have the static locational reading, paraphrasable as in case (24)(ii).

(24) The child jumped on the bed.

(i) The child jumped onto the bed.

(ii) The child was on the bed jumping (on it).

With the strict requirement that the adjunct phrases be of type PLACE, a non-canonical sentence such as (25) will cause a type mismatch failure at the IL composition step.[36] The PP *through the tunnel* will be syntactically parsed as an adjunct because the lexical syntactic form for the verb *sleep* has no arguments and so, from the parser's point of view, it must be attached as an adjunct. When the IL composition step attempts to attach the PATH-typed IL form for this adjunct PP to the PLACE-typed adjunct position in *sleep*'s IL form, a mismatch is detected. Indeed we find that, in German, this sentence cannot be translated literally. The spatial sense of *durch,* a German equivalent of *through,* is not acceptable as head of an adjunct phrase and so an alternate paraphrase is required.

(25) (i) English: She slept through the tunnel.
(ii) German: Sie schlief waehrend sie durch den Tunnel ging.
'She slept while she through the tunnel went'

Operationally, the composition fails because the intransitive verb has no argument position and the PP "falls" into the adjunct position. But the general question concerns how to treat PATH IL forms when they appear in adjunct PLACE-typed positions. Note that sentence (25)(i) is indeed grammatical and needs to be interpreted by the MT system. Furthermore, following our principle for division of labor, we have sought a solution to the processing of this non-canonical sentence that reflects the logical inferencing required for its interpretation. We have been guided in our solution to interpreting this sentence both by our IL syntax that permits factoring a verb's meaning into a predicator with a constant modifier and by the German translation of the sentence. Given a durative verb, i.e., one whose KR concept refers to a situation that lasts over some period of time, when encountering a PATH-typed adjunct spatial relation, the MT system may *augment* the interpretation of the sentence in order to recover the logically inferable spatial motion (here introduced in English by the PP). In particular, the subject's IL form and the PP's IL form are composed into a newly introduced $\mathrm{GO}_{\mathrm{Loc}}$ general spatial motion IL form while the constant in the verb's IL form is attached as a modifier of the $\mathrm{GO}_{\mathrm{Loc}}$ predicate.

The analysis here for sentence (25)(i) contrasts with that of the sentence *the truck rumbled into the yard* given earlier in (4). In both cases a directional PP appears following an intransitive verb (i.e., whose lexical syntactic form has no argument position available at which to attach the PP). The difference between the *sleep* and the *rumble* analyses is as follows. In the latter case we argue that the semantics of the verb's event entitles us to encode, as *lexically implicit,* both the sound and the motion within the single lexical form: the rumbling sound and the rumbling motion are necessarily concommitent within the same event. By contrast, in the former case it is the KR-based properties of the verb's event that entitles us to *logically infer* when it is compatible with a concurrent motion event. We are able to capture this semantic distinction in the MT system by virtue of the separate levels of representation.

[36] This sentence often causes readers, on a first reading, to pause. Had the sentence been about sleeping through the night, i.e., had the word *through* been used in a strictly temporal sense rather than as in (25)(i), the reading would have been unproblematic. It appears, though we have seen no psycholinguistic research on this, that the *through* phrase in sentence (25)(i) requires some extra processing by readers and they are aware of this.

5.3 During IL Decomposition

Once a full IL form for the SL input sentence has been analyzed and checked in the filtering phase, the MT system begins the first step in generating the TL output by determining the range of TL lexical IL forms that can "cover" the full IL form. In our approach using predicate decompositional structures for the IL level of representation, the full IL form is the guiding structure out of which lexical IL forms are extracted or decomposed. The lexical IL forms must then be found in the TL lexicon and the associated TL word (or phrase) can be retrieved. In this section we discuss our work building an indexing data structure for organizing the MT system's lexical IL forms and retrieving a TL entry given such a form.

5.3.1 IL lexicon: An Example

As an example of our approach, consider a small portion of an IL lexicon, as shown in Figure 6, that consists of seven lexical-semantic forms (RLCS's without language-specific annotations) and their respective lexicalizations in English and German. (We have placed '***' in the figure where no word for the particular language exactly matches the forms listed.) The German words are *veranlassen, bewegen, transportieren,* and *fahren* and the English words are *cause, move, transport,* and *bus*. Prior to translation runtime, the lexical entries are classified and the IL Lexicon in Figure 6 is produced. Note that, if we were to classify the three verbs used earlier in (14) as part of this structure in Figure 6, *lift* would be aligned with [CAUSE(_ ,[GO$_{\text{Loc}}$(_ , _)])] and *ascend* and *climb* would be aligned with [GO$_{\text{Loc}}$(_ , _)].

The organization of the IL is guided by the syntax of the lexical IL forms held in the nodes of the data structure itself. Our goal here is to be able both to delimit the space of lexical IL forms available in the MT system(what was referred to above as the lexicalization space) and to efficiently retrieve lexical IL forms that correspond to the substructures found within a lexical IL form.

With respect to the first goal of delimiting the lexicalization space, we have found this is critical to developing RLCSs for a new language to be added to the MT system. When a non-native speaker of English begins building the RLCSs for lexical items in their native language, difficulties arise immediately as they attempt to understand the IL predicators that are given English labels. By having the IL predicators ground in the KR ontology, the deeper conceptual knowledge stored in the corresponding KR concept is also available as another source in defining the basic IL terms.

The second goal of retrieving lexicalized substructures for a given IL form is motivated by the decomposition task and the specific way in which the IL forms are constructed and lexicalized. Here we will examine the structures of a few causative verbs and their non-causative "counterparts" to clarify this relation between the IL structures and the decomposition task.

Consider the verb *wash* in sentences (26).

(26) (i) The storm washed the ship ashore.

(ii) The ship washed ashore.

The English verb *wash* has a causative reading in (26)(i) that is also conveyed in the paraphrase, the storm caused the ship's being washed ashore. *Wash* also has a non-causative

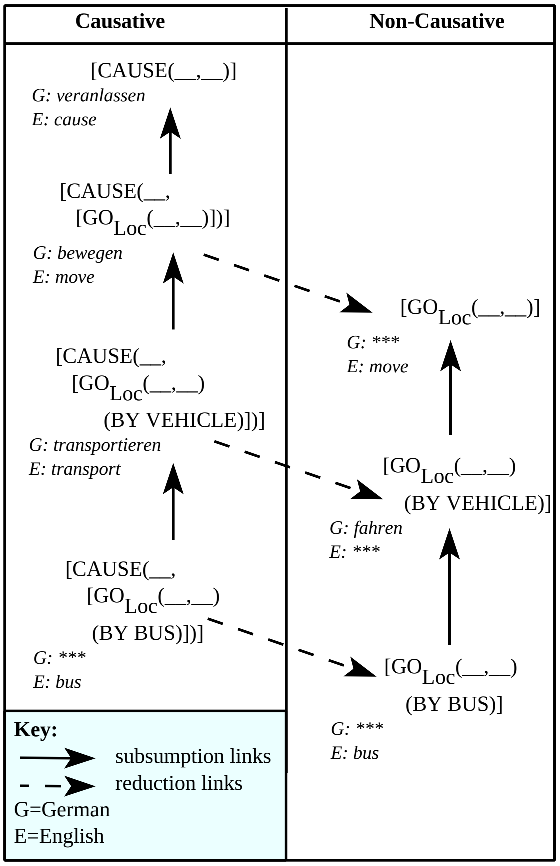

Figure 6: RLCS's with Directed Subsumption Links and Crossover Reduction Links in Semantic Ontology

reading, as in (26)(ii). From this minimal pair, it is clear that the lexical semantics of this non-causative sense is contained within the lexical semantics of the causative sense. In the predicate decompositional form of our IL, this semantic containment relationship is captured by having the lexical IL form of the non-causative sense as a substructure in the lexical IL form of the causative sense.

Now consider this relation as represented in the IL lexicon: a *reduction* link connects the lexical IL form of a causative form to its non-causative counterpart. This type of link is used in the IL decomposition process when no causative form can be found in the TL lexicon. By following a reduction link, the search algorithm in the decomposition process will seek an alternate translation using the non-causative form together with a lexicalization for the cause predicate. In figure 6, the reduction links (dashed arrows) run from left to right, connecting individual entries in the column of lexical IL forms for causative verbs with their non-causative counterparts.

Figure 6 also contains a second type of link relevant to the IL decomposition process, *subsumption* links shown by solid arrows. For example, the IL form for the English word *move* (in the left-hand column, this refers to externally caused motion) stands in a subsumption relation with (i.e., is more specific than) the English word *cause*. On the other hand, the RLCS for the English word *move* (left-hand column) stands in a reduction relation with its non-causative English counterpart *move* (right-hand column, this refers to motion that is not externally caused or whose actual cause is linguistically unspecified). This latter link is non-standard as KR schemes go—in fact, the left-hand *move* is neither more specific nor more general than the right-hand *move*; rather, there is the linguistically-relevant structural relation between the causative and non-causative IL forms of the same verb that is logically distinct from subsumption relations. The causative IL form will contain substructures corresponding to its causing and its resulting subevents. Thus, the reduction links serve to partition verbal meaning into subcomponents where each part has its meaning preserved. The subsumption links lead to structures that have a narrower, more specific sense than their subsumers and that bear a well-defined structural relation to those of their subsumers. While these links do not preserve meaning subcomponents, by virtue of their structural relation, they provide another search path for the decomposition algorithm to follow and an alternative resource in lexicalizing an IL form.

Recall from Section 2.3 that there are three mismatch cases between SL and TL words that we are attempting to handle: subsumes, subsumed-by, and overlapping.[37] Consider the case of translating from English to German. In Figure 6 in the RHS column, we can see first that the non-causative English verb *bus* has no German lexical equivalent. However, since German does have a slightly more general non-causative verb *fahren,* the translation algorithm can opt for a *subsumes* relation to resolve the mismatch and select *fahren* as the

[37] We note briefly here that without a mismatch, the runtime processing is straightforward. Consider the translation of the German word *bewegen* into the English word *move.* First the RLCS for the German word is retrieved and composed with other sentence elements to form the full IL form in the analysis phase. Then, during generation, the same term classifier that placed RLCSs (lexical IL forms) for English and German entries into the IL lexicon pre-runtime, now determines where the relevant portion of the full IL form falls among the English RLCSs in the IL lexicon. Since *bewegen* is synonymous with one sense of the word *move,* the classifier will match that portion of the composed IL form with the RLCS entry for that synonymous sense; this will then be realized in English as *move.*

translation of *bus.* Suffice it to say here that the selected verb, *fahren* becomes the head of the TL phrase and the information "dropped" in the move up the ontological hierarchy (from non-causative *bus* to *fahren*) becomes the restrictive modifier phrase to that head. (The information for the phrase is readily identifiable from a structural comparison between subsumer and subsumee.) This procedure is the verb phrase analog to the approach taken by (Sondheimer, Cumming, and Albano, 1990) with noun phrases. The translations for the sentences in (27) show the result of this procedure. As with *bus,* translating the non-causative verb *train* to German also requires using a combination of partial matches with German words to achieve the full coverage matching of the English verb's IL form: a subsuming verb becomes the main verb, *fahren,* and a modifier phrase, *mit dem Zug,* restricts the sense of the more general subsuming verb.[38]

(27) (i) They bused into town.

(ii) Sie fuhren mit dem Autobus in die Stadt.
'They *vehicled* by bus into town'

(28) (i) They trained into town.

(ii) Sie fuhren mit dem Zug in die Stadt.
'They *vehicled* by train into town'

A *subsumes* relation is not always available to resolve lexicalization gap mismatches. The non-causative English verbs *go* and *move* are very general and do not exist as a simple lexical entry in German.[39] At that level in the ontology, there is no more general non-causative concept to tap for the translation. For example, in sentence (29), the common English verb *go* is translated into the German equivalent of *run, laufen,* a verb whose sense is subsumed by that of *go.* That is, in this case the selection algorithm must opt for a *subsumes-by* relation by translating the SL word as a TL word that is lower down in the ontology. In translating verbs, this selection depends crucially on finding a TL verb whose constraints are met by both the SL verb and its arguments.

(29) (i) Because of the dog, the cat went away.

(ii) Wegen dem Hund lief die Katze weg.
'Because of the dog, ran the cat away'

Finally an *overlap* relation occurs in translating the causative English verb *bus* where again we find no corresponding German lexical entry in the ontology, as in sentence (30). One option needed for translating this verb in a formal style of speech (such as in a legal document) involves decomposing its meaning into the comparable phrase *to cause to go by bus.* While the *cause* concept (for *veranlassen*) subsumes that of the causative *bus,* a

[38] Not all speakers of English accept verbs of this form. As with all acceptability judgements, we find variation among native speakers. Once a word such as *bus* is listed in the dictionary (here Webster's Ninth), as a transitive verb along with both definitions "to travel by bus" and "to transport by bus", i.e., precisely the noncausative and causative readings, then we treat it as acceptable.

[39] German *sich bewegen* is the closest, non-simple equivalent to the non-causative English *move.*

reduction link relation (see lowest dashed arrow in Figure 6)—not a subsumption relation – is needed to capture *go by bus*.[40]

(30) (i) The man bussed the animals to the zoo.

(ii) Der Mann veranlasste, dass die Tiere mit den Bus zum Zoo gefahren wurden.
'The man caused/made (that) the animals by bus to the zoo *vehicled* were'

5.3.2 Sketch of Lexical Selection Algorithm

We now describe our lexical selection algorithm. The role of this algorithm is to return the target language (TL) lexical entries whose IL forms cover the IL form submitted as input.[41] A key point about the algorithm is that the classifier, which is used prior to processing time, is also a main driving component of the lexical selection scheme in the Lookup step.

The details are given here:

- **Input IL Form Lookup:** Search the IL lexicon on the basis of the algorithm's input: a composed IL form (i.e., the CLCS of (Dorr, 1993)) or part thereof, in order to determine its hierarchical position. If no corresponding form is found, return to calling program (input is not a syntactically valid *lexical* IL form).
- **RLCS Retrieval:** Retrieve, from the TL lexicon, instances of TL words indexed from the current entry, i.e., IL lexicon entry in current hierarchical position.
- **Match Detection:** If there exists one or more TL RLCS for the current entry, return the associated TL lexical items(s) and entry form.
- **Overlapping Case:** In the absence of an exact match, first follow a reduction link, if available,[42] to a substructure entry and then follow subsumption link to a root entry.[43] If the substructure entry has TL RLCS, return the root and substructure TL RLCSs and entries. If the substructure entry does not have a TL RLCS, return the root RLCS and entry and then recursively call this algorithm with the substructure entry as input.
- **Familial Information Return:** In the absence of structural coverage (the match or overlap cases above), return, where available, the ontological parent (subsumer) and list of ontological children (subsumed-by) and current entry.

Note that the overlap reduction has a higher standing than subsumption. This is because our research goal has been to experiment with retrieving the fullest possible exact structural coverage of the IL input; if we attempt subsumption before reduction, the calling process

[40] The careful reader will note both within a language and cross-linguistically, that in general, when paraphrasing a single verb with a periphrastic, or multiple word expression, some subtle event-level changes in meaning will occur.

[41] The task of identifying the input IL form is based on the fully composed IL form for the SL input and is carried out by a lexical options generator that calls this algorithm.

[42] If no reduction link is available, the overlap case does not apply.

[43] 'Substructure entry' refers to the IL lexicon entry at the end of the reduction link and 'root entry' refers to the IL lexicon entry for the root predicate of the current entry. All root predicate entries in the IL lexicon have TL RLCSs.

must derive the structural difference between the IL forms of the subsumer or subsumee and the current entry, and then generate the modifying phrase that best approximates a structural cover—and the reduction option will never be reached. Nothing in principal however preempts revising the algorithm to run both the reduction and subsumption options. This adds to the set of lexicalization options available at the next processing stage in the generation phase.

5.3.3 Implementation and Related Approaches

We have implemented this approach in LOOM (MacGregor, 1991), a KL-ONE-like term classifier and its concept definitions. (Other details are given in (Dorr et al., 1994).) LOOM and other frame-based systems (e.g., KL-ONE and KRL) have also been used by a number of other researchers, including (Brachman and Schmolze, 1985; MacGregor, 1991; Woods and Brachman, 1978). As mentioned above, our goal in developing the IL lexicon in LOOM was to test the feasibility of having this data structure within a KR system with a classifier where the IL primitives in the IL lexicon entries would also be established as concepts in the KR system's ontology, maintaining the consistency in the grounded terms at the IL/KR interface. As part of this feasibility experiment, we built only a small KR ontology for the purposes of supporting the primitives in the IL lexicon and the inferences in the composition and filtering phases.

Our approach is similar to that of DiMarco, Hirst, and Stede (1993), which also uses the LOOM classifier, but with the complementary goal of handling fine-grained, stylistic variations. There are also similarities between our approach and that of (Nirenburg et al., 1992), in that we seek a single unifying data structure to establish a range of semantic relations among words.[44] Because our primary focus is on the development of the IL and the IL lexicon, our KR entries are driven by our lexicalization needs and the IL primitives in the IL syntax. (By contrast, in many large-scale KBMT projects, the KB concepts in the large ontologies drive the definitions of lexical entries.)

Alternative KR formalisms have been explored by a number of researchers including (Wu and Palmer, 1994; Palmer and Wu, 1995; Ali and Shapiro, 1993; Iwanska, 1993; Quantz and Schmitz, 1994; Schubert, 1991; Sowa, 1991). In the approach of (Wu and Palmer, 1994; Palmer and Wu, 1995), lexical mismatches are resolved through the use of a conceptual lattice based on an adapted version of (Levin, 1993). While this approach is similar in spirit to our method of mismatch resolution, Palmer and Wu propose an *approximate* matching scheme based on "semantic closeness" as determined by numerical distances between semantic classes, whereas our scheme relies on a full *structural coverage* matching as determined by alignment among lexical-semantic components that systematically map into syntactic structure. We argued in Section 3, that we must necessarily preserve substructure information (i.e., LCS descriptions) in order to provide an adequate translation into a TL surface structure; thus, a numerical measure of semantic closeness would not be adequate for our purposes. Moreover, the *structural coverage* matching scheme is a prerequisite for recursive

[44] Strictly speaking, the lexicons in interlingua-based MT systems are not restricted to word-level entries. For the purposes of this paper however we will refer to "words" in the lexicons, setting aside the details about other types of lexical entries. See (Levin and Nirenburg, 1993) for further discussion on extending the range of lexical entries in MT systems.

instantiation of substructures that are not necessarily aligned with the full structure of the IL during the translation process. We have already seen examples of such cases in Section 5.3.1. Finally, the classification system of Levin (1993) has been refined in later work (Levin and Rappaport-Hovav, 1995a; Levin and Rappaport-Hovav, 1995b), where regularities in lexical representations—and the (parameterized) rules that operate on these representations—can be exploited for a more economical approach to MT. We saw several cases where we could benefit from this economy in Section 4. Such regularities would be lost if we were to adopt a numerical "closest match" scheme.

Other approaches have different objectives or address deeper conceptual issues. One example out of many is the work of (Iwanska, 1993), which is concerned with notions such as logical inferences, entailment, negation, and quantification. The primary concern in Iwanska's work is the population of a knowledge base and the provision of a framework for truth maintenance and queries. While this work has certain elements in common with our approach (e.g., the representation of entailment, which is similar to our notion of classification), the framework is more applicable to a discourse analysis system than to the problem of mismatch handling in MT.

6 Conclusions

We have described a non-uniform approach to interlingual MT, in which distinct representational languages are used for different types of knowledge. The key contribution of this work is that it provides a linguistically-motivated "division of labor" across multiple representation levels. We have shown that this multi-representation approach has not complicated, but rather has readily facilitated, the identification and construction of systematic relations at the interface between each level.

The novelty of our approach is that the IL representation is designed to preserve the "semantic structure" of the SL and TL sentences while still retaining the "semantic content" by means of conceptually grounded primitives. The preservation of semantic structure allows us to readily map between the IL and the syntax; the preservation of semantic content allows us to readily map between the IL and KR. Approaches that ignore the structural nature of lexical items and their combination stand to lose the benefit of regularities that exist *within* a language (in the lexicon) as well *across* languages (in the surface syntactic form). By decomposing verbs into *predicate* structure and non-predicate *constants*, we exploit such regularities.

We have cited examples from the spatial domain in support of our non-uniform representational approach, and we have demonstrated the cross-linguistic applicability of the multi-level MT design in this domain. We have implemented this design in the PRINCITRAN system, which handles a wide range of cross-linguistic mismatches. Areas of future investigation include development of a framework for handling multiple sentences; in particular, a rich theory of context coupled with our rules of composition might allow us to identify cases where the IL underlying the SL sentence must be either reduced or augmented in order to produce an appropriate TL sentence.

References

Ali, S. and S. Shapiro. 1993. Natural Language Processing Using a Propositional Semantic Network with Structured Variables. *International Journal of Minds and Machines, Special Issue on Knowledge Representation for Natural Language*, 3:421–451.

Anderson, J. 1971. *The Grammar of Case: Towards a Localist Theory.* Cambridge University Press, Cambridge, England.

Barnett, J., I. Mani, P. Martin, and E. Rich. 1994. Reversible Machine Translation: What to Do When the Languages Don't Match Up. In T. Strzalkowski, editor, *Reversible Grammar in Natural Language Processing.* Kluwer Academic Publisher.

Barwise, J. and J. Perry. 1983. *Situations and Attitudes.* The MIT Press, Cambridge, MA.

Beaven, J. 1992. Shake and Bake Machine Translation. In *Proceedings of Fourteenth International Conference on Computational Linguistics*, pages 603–609, Nantes, France.

Brachman, R. J. and J. Schmolze. 1985. An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, 9(2):171–216.

DiMarco, C., G. Hirst, and M. Stede. 1993. The Semantic and Stylistic Differentiation of Synonyms and Near-Synonyms. In *Working Notes for the AAAI Spring Symposium on Building Lexicons for Machine Translation*, Technical Report SS-93-02, pages 114–121, Stanford University, CA.

Dorr, B. 1993. *Machine Translation: A View from the Lexicon.* The MIT Press, Cambridge, MA.

Dorr, B., J. Garman, and A. Weinberg. 1995. From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT. *Machine Translation*, 9(3).

Dorr, B., D. Lin, J. Lee, and S. Suh. 1995. Efficient Parsing for Korean and English: A Parameterized Message Passing Approach. *Computational Linguistics*, 21:2:255–263.

Dorr, B. and C. Voss. 1993. Machine Translation of Spatial Expressions: Defining the Relation between an Interlingua and a Knowledge Representation System. In *Proceedings of Twelfth Conference of the American Association for Artificial Intelligence*, pages 374–379, Washington, DC.

Dorr, B., C. Voss, E. Peterson, and M. Kiker. 1994. Concept Based Lexical Selection. In *AAAI 1994 Fall Symposium on Knowledge Representation for Natural Language Processing in Implemented Systems*, New Orleans, LA.

Dorr, B. J. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633.

Dorr, B.J., J. Hendler, S. Blanksteen, and B. Migdalof. 1995. Use of LCS and Discourse for Intelligent Tutoring: On Beyond Syntax. In M. Holland and J. Kaplan and M. Sams, editor, *Intelligent Language Tutors: Balancing Theory and Technology.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Grimshaw, J. 1994. Semantic Structure and Semantic Content in Lexical Representation. unpublished ms., Rutgers University, New Brunswick, NJ.

Heine, B., U. Claudi, and F. Huennemeyer. 1991. *Grammaticalization: A Conceptual Framework.* University of Chicago Press, Chicago, IL.

Iwanska, L. 1993. Logical Reasoning in Natural Language: It is all about Knowledge. *International Journal of Minds and Machines, Special Issue on Knowledge Representation for Natural Language*, 3:475–510.

Jackendoff, R. 1983. *Language and Cognition.* The MIT Press, Cambridge, MA.

Jackendoff, R. 1990. *Semantic Structures.* The MIT Press, Cambridge, MA.

Jackendoff, R. 1991. Parts and Boundaries. In B. Levin and S. Pinker, editors, *Lexical and Conceptual Semantics.* Blackwell Publishers, Cambridge, MA.

Jackendoff, R. 1992. What is Semantic Structures About? *Computational Linguistics*, 18:2:240–242.

Kameyama, M., R. Ochitani, S. Peters, and H. Sirai. 1991. Resolving Translation Mismatches with Information Flow. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 193–200, University of California, Berkeley, CA.

Kamp, H. 1981. A Theory of Truth and Semantic Representation. In J. A. G. Groenendijk, U. M. V. Janssen, and M. J. B. Stokhof, editors, *Formal Methods in the Study of Language*, volume 136. Mathematical Centre Tracts, Amsterday, The Netherlands, pages 277–322.

Kay, M., J. M. Gawron, and P. Norvig. 1994. *Verbmobil: A Translation System for Face-to-Face Dialog*, volume 33. CSLI Lecture Notes, Stanford, CA.

Kinoshita, S., J. Phillips, and J. Tsujii. 1992. Interaction Between Structural Changes in Machine Translation. In *Proceedings of Fourteenth International Conference on Computational Linguistics*, pages 679–685, Nantes, France.

Langacker, R. 1987. *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites.* Stanford University Press, Stanford, CA.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, IL.

Levin, B. and M. Rappaport-Hovav. 1995a. The Elasticity of Verb Meaning. In *Proceedings of the Tenth Annual Conference of the Israel Association for Theoretical Linguistics and the Workshop on the Syntax-Semantics Interface*, Bar Ilan University, Israel.

Levin, B. and M. Rappaport-Hovav. 1995b. *Unaccusativity: At the Syntax-Semantics Interface.* The MIT Press, Cambridge, MA.

Levin, L. and S. Nirenburg. 1994. The Correct Place Of Lexical Semantics in Interlingual MT. In *Proceedings of Fifteenth International Conference on Computational Linguistics*, Kyoto, Japan.

Levin, Lori and Sergei Nirenburg. 1993. Principles and Idiosyncracies in MT Lexicons. In *Working Notes for the AAAI Spring Symposium on Building Lexicons for Machine Translation*, pages 122–131, Stanford University, CA.

Lindop, J. and J. Tsujii. 1991. Complex Transfer in MT: A Survey of Examples. Technical report ccl/umist report 91/5, Center for Computational Linguistics, UMIST, Manchester, UK.

Lytinen, S. and R.C. Schank. 1982. Representation and Translation. Technical Report Technical Report 234, Department of Computer Science, Yale University, New Haven, CT.

MacGregor, R. 1991. The Evolving Technology of Classification-Based Knowledge Representation Systems. In J. Sowa, editor, *Principles of Semantic Networks.* Morgan Kaufmann, San Mateo, CA.

Mani, I. 1995. An Integrative, Layered Approach to Lexical Semantics and Its Application to Machine Translation. In *Working Notes for the AAAI Spring Symposium on Representation and Aquisition of lexical Knowledge: Polysemy, Ambiguity, and Generativity*, Technical Report SS-95, Stanford University, CA.

Melby, A.K. 1986. Lexical Transfer: Missing Element in Linguistic Theories. In *Proceedings of Eleventh International Conference on Computational Linguistics*, Bonn, Germany.

Nirenburg, S., J. Carbonell, M. Tomita, and K. Goodman. 1992. *Machine Translation: A Knowledge-Based Approach.* Morgan Kaufmann, San Mateo, CA.

Nirenburg, S. and I. Nirenburg. 1988. A Framework for Lexical Selection in Natural Language Generation. In *Proceedings of Twelveth International Conference on Computational Linguistics*, pages 471–475, Budapest, Hungary.

Palmer, M. and Z. Wu. 1995. Verb Semantics for English-Chinese Translation. *Machine Translation*, 9(4).

Pugeault, F., P. Saint-Dizier, and M.G. Monteil. 1994. Knowledge Extraction From Texts: A Method For Extracting Predicate-Argument Structures From Texts. In *Proceedings of Fifteenth International Conference on Computational Linguistics*, Kyoto, Japan.

Quantz, J.J. and B. Schmitz. 1994. Knowledge-Based Disambiguation for Machine Translation. *International Journal of Minds and Machines, Special Issue on Knowledge Representation for Natural Language*, 4:39–57.

Rosetta, M. T. 1994. *Compositional Translation.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

Rupp, C. J., R. Johnson, and M. Rosner. 1992. Situation Schemata and Linguistic Representation. In M. Rosner and R. Johnson, editors, *Computational Linguistics and Formal Semantics.* Cambridge University Press, Cambridge, England, pages 191–221.

Schank, R.C. 1973. Identification of Conceptualizations Underlying Natural Language. In R.C. Schank and K.M. Colby, editor, *Computer Models of Thought and Language.* Freeman, San Francisco, CA, pages 187–247.

Schank, R.C., editor. 1975. *Conceptual Information Processing.* Elsevier Science Publishers, Amsterdam, Holland.

Schank, R.C. and R. Abelson, editors. 1977. *Scripts, Plans, Goals, and Understanding.* Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Schubert, L. 1991. Semantic Nets are in the Eye of the Beholder. In J. Sowa, editor, *Principles of Semantic Networks.* Morgan Kaufmann, San Mateo, CA.

Slobin, D. forthcoming. Typology and Rhetoric: Verbs of Motion in English and Spanish. In M. Shibatani and S. Thompson, editors, *Grammatical Constructions: Their Form and Meaning.* Oxford University Press.

Sondheimer, N., S. Cumming, and R. Albano. 1990. How to Realize a Concept: Lexical Selection and the Conceptual Network in Text Generation. *Machine Translation*, 5(1):57–78, March.

Sowa, J. 1991. Toward the Expressive Power of Natural Language. In J. Sowa, editor, *Principles of Semantic Networks.* Morgan Kaufmann, San Mateo, CA, pages 157–189.

Talmy, L. 1983. How Language Structures Space. In H.L. Pick and L.P. Acredolo, editors, *Spatial Orientation: Theory, Research, and Application.* Plenum Press, New York, pages 225–282.

Talmy, L. 1985. Lexicalization Patterns: Semantic Structure in Lexical Forms. In T. Shopen, editor, *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon.* University Press, Cambridge, England, pages 57–149.

Verrière, G. 1994. Manuel d'utilisation de la structure lexicale conceptuelle (LCS) pour représenter des phrases en français. Research note, IRIT, Université Paul Sabatier, Toulouse, France, June.

Voss, C., B. Dorr, and M. Ülkü Şencan. 1995. Lexical Allocation in Interlingua-Based Machine Translation of Spatial Expressions. In *Working Notes for IJCAI-95 Workshop on the Representation and Processing of Spatial Expressions*, Montreal, Canada.

Whitelock, P. 1992. Shake-and-Bake Translation. In *Proceedings of Fourteenth International Conference on Computational Linguistics*, pages 784–791, Nantes, France.

Woods, W.A. and R.J. Brachman. 1978. Research in Natural Language Understanding. Quarterly technical report progress report no. 1, Bolt, Beranek, and Newman, Cambridge, MA.

Wu, Z. and M. Palmer. 1994. Verb Semantics and Lexical Selection. In *Proceedings of Association for Computational Linguistics*, Las Cruces, NM.

Zwarts, J. and H. Verkuyl. 1994. An Algebra of Conceptual Structure: An Investigation Into Jackendoff's Conceptual Semantics. *Linguistics and Philosophy*, 17:1–24.