Concept-Based Lexical Selection*

Bonnie J. Dorr and **Clare R. Voss**
Department of Computer Science
A.V. Williams Building
University of Maryland
College Park, MD 20742
e-mail: {bonnie,voss}@cs.umd.edu

Eric Peterson and **Michael Kiker**
The MITRE Corporation
7525 Colshire Dr.
McLean, VA 22102-3481
e-mail: {eric,mjkiker}@starbase.mitre.org

Abstract

We present a concept-based approach to the problem of lexical selection which allows us to deal precisely with lexicalization gaps or mismatches in mapping from a source language into a target language. We adopt a linguistically motivated scheme that makes use of a decompositional representation which can be classified in terms of a KR ontology. A main contribution of our representation scheme is the ability to handle cases where two concepts overlap, but neither subsumes the other. We have demonstrated the utility of our approach for mismatched verbs in English and German.

1 Introduction

This paper presents a concept-based approach to the problem of lexical selection, i.e., the task of choosing an appropriate target-language term in generating text from an underlying meaning representation. Our goal is to provide a linguistically motivated scheme for handling lexical mismatches; we view this not solely as a language-to-language problem in the domain of machine translation, but as the larger problem of handling "lexicalization gaps" in the mapping from the knowledge base to the surface realization.

We describe an architecture for lexical selection in interlingua-based machine translation (MT) systems using KL-ONE-like concepts [Brachman and Schmolze, 1985] for grounding the lexical semantic descriptions of both target-language (TL) and source-language (SL) words. The concepts are classified into a semantic ontology for each supported language using LOOM, a KL-ONE-like term classifier. We believe this architecture is an advance over previous designs for three reasons: (*i*) It provides a more precise method for identifying and resolving mismatches between SL and TL words; (*ii*) It facilitates a graceful combination of the knowledge base (KB) and the lexical-semantic representation; and (*iii*) It offers greatly increased semantic expressive power, enabling the representation of complex constraints and relationships between intra-concept constituents.[1]

With respect to representing word meaning, two camps in MT research have prevailed: (*i*) lexical conceptual structures (LCS), in the spirit of [Jackendoff, 1983]; and (*ii*) frame-based systems in the spirit of [Nirenburg *et al.*, 1992]. We combine the most promising aspects of both approaches, i.e., the minimalism and decompositionality of LCSs,[2] as well as the ability to perform frame-based reasoning in an object-oriented environment. We take the view that, as semantic lexicographers, we can benefit from the techniques used by both LCS and knowledge representation (KR) experts in designing semantic representations of word meaning. Specifically, we use LOOM [MacGregor, 1991], a KL-ONE-like term classifier, and its concept definitions — reasoning frames with the added power of logical constraints.[3]

Our approach is interlingual (i.e., language independent); our goal is to demonstrate the feasibility of encoding a lexical-semantic representation by means of concepts in a KB, thus allowing the interlingua (IL) to capture aspects of both types of information. We feel that

*Bonnie Dorr and Clare Voss are supported, in part, by a National Science Foundation Young Investigator Award IRI-9357731 and LOGOS Corporation, by the Army Research Office under contract DAAL03-91-C-0034 through Battelle Corporation, and by the Army Research Institute under contract MDA-903-92-R-0035 through Microelectronics and Design. We wish to thank an anonymous reviewer for guidance and commentary leading to the final version of this paper. We also thank Birthe Stoll at the Freie Universität Berlin for the German translations she provided.

[1] Intra-concept constituents are stored within the concept roles and are constrained by well-formedness conditions.

[2] We use the term "decompositionality" to mean the representation of concepts as combinations of atomic units of meaning.

[3] Barnett, Mani, and Rich [Barnett *et al.*, 1994, p. 363] suggest a similar use of a term classifier.

the proposed architecture not only meets this original goal, but is a decisive foundation for increasing the expressive power of the lexical semantic representation and hence the precision of the surface realization. One important result is the ability to deal precisely with the case where the target language lacks a word that directly matches the IL fragment to be generated in the target language.

We should point out that although the focus of this technique is on machine translation (MT), it would work as well for knowledge-based language generation.[4]

2 A Brief Architectural Description

Like [Nirenburg *et al.*, 1992], we seek to determine a range of semantic relations among words[5] in our semantic ontology. Our approach differs, however, in that it uses a term classifier to set up an ontology of lexical entries for each supported language. For a given target language, its ontology helps to determine, through classification, the lexical entry which most closely matches the portion of the IL form being generated.

Before translation runtime, the classifier is employed to form separate lexical entry ontologies for each supported language. At translation runtime, the lexical selection process attempts to realize a given portion of an IL (CLCS)[6] by running the term classifier on the relevant IL structure and then determining where the concept falls in the ontological hierarchy.[7]

Consider an ultra-minimal semantic ontology, shown in Figure 1, that consists of seven lexical-semantic forms (RLCS's) 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 word descriptions are classified and the ontology in Figure 1 is produced.

We now consider the runtime process of lexical selection during the translation of the German word *bewegen* into the English word *move.* First, the RLCS for

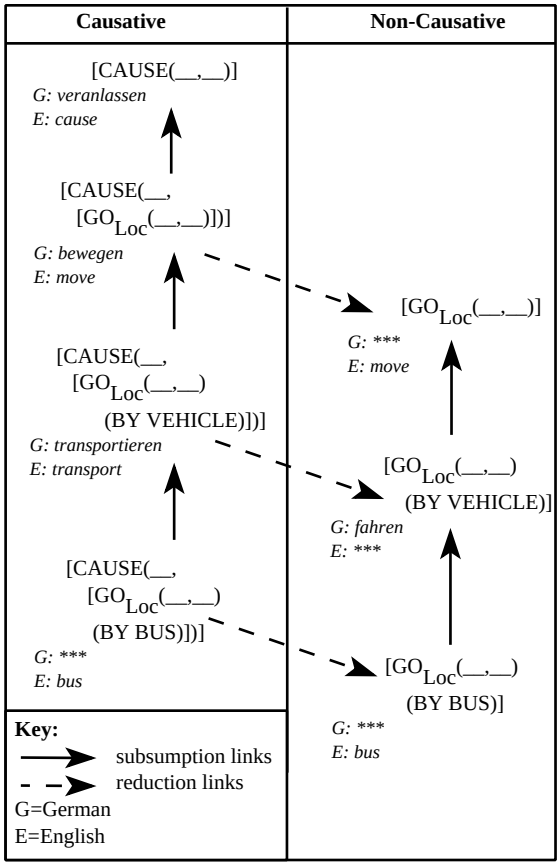

Figure 1: RLCS's with Directed Subsumption Links and Crossover Reduction Links in Semantic Ontology

bewegen is retrieved and composed with other sentence elements to form the CLCS in the analysis phase of the translation. Then in the generation phase as part of the algorithm for selecting a TL word, the same term classifier that was used to build the English lexicon is re-used to determine where the relevant portion of the CLCS falls in the ontology. Since *bewegen* is synonymous with one sense of the word *move*, the classifier will find that this portion of the CLCS matches the ontology entry for the causative sense of the word *move;* this will then be realized in English as *move*.

This example has been kept simple for expository reasons, but section 5 outlines the more complicated cases, i.e. where the matching fails.

3 Lexical Mismatches

We now discuss cases of mismatch between the source and target languages. The mismatch problem in MT has received increasingly greater attention in recent literature (see [Dorr, 1994], [Dorr and Voss, 1993], [Barnett *et al.*, 1994], [Beaven, 1992], [Kameyama *et al.*, 1991], [Kinoshita *et al.*, 1992], [Lindop and Tsujii, 1991], and [Whitelock, 1992] as well as related discussion in [Melby, 1986] and [Nirenburg and Nirenburg, 1988].) In particular, [Barnett *et al.*, 1994] divide distinctions between the source language and the target language into two cate-

[4]See [Stede, 1993] for further discussion with respect to multilingual generation from a knowledge base.

[5]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.

[6]Composed Lexical Conceptual Structure or instantiated Root Lexical Conceptual Structure (RLCS). RLCS's are the actual semantic concept descriptions for the words in a lexicon. The terms RLCS and CLCS are borrowed from [Dorr, 1993, Dorr, 1994].

[7]In this paper we focus on the lexical selection component of the generation process. That is, we address the problem of selecting the appropriate TL word after the CLCS has been partitioned into relevant components for structural realization. See work by [Dorr *et al.*, 1994b] on a message-passing approach to syntactic analysis which is currently being extended to handle the structural-realization component of the generation process.

gories: translation *divergences*, in which the same information is conveyed in the source and target texts, but the structures of the sentences are different (as in previous work by [Dorr, 1994]); and translation *mismatches*, in which the information 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 ignored 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]) 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 of the IL to resolve the divergence problem. We focus on the problem of mismatches for the remainder of this paper; the reader is referred to [Dorr, 1994] for an in-depth treatment of the divergence problem.

Our solution to the mismatch problem involves the use of the LOOM KL-ONE-like term classifier. The approach is similar to that of [DiMarco *et al.*, 1993], which also uses the LOOM classifier, but with the complementary goal of handling fine-grained, stylistic variations. 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], among others. Alternative KR formalisms have been explored by a number of researchers including [Shapiro, 1993, Iwanska, 1993, Quantz and Schmitz, 1994, Schubert, 1991, Sowa, 1991]. Most of these approaches have different objectives or address deeper conceptual issues. An example 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 lexical selection in a generation system.

The remainder of this section scopes out the space of possible lexical mismatches and describes our specific focus within this space.

3.1 Space of Lexical Mismatches

We have found that, although the problem of lexical mismatches has been characterized in a number of ways within the MT research literature, there does not exist an agreed-upon partitioning of the data in this *problem space*. Here we first take a brief look at a classic case of a lexical mismatch and a few ways of dividing up the problem space. We then present our own alternative approach to identifying mismatches.

Consider one classic mismatch case: Spanish has two words *pez* and *pescado* that correspond to the English word *fish. Pez* is used for a default generic, or unmarked, situation when it is assumed to include all fish,[8] and *pescado* is used in the specific, marked situation for a fish that has been caught and is no longer in its natural state.

The simplest **descriptive** characterization of this example is that it is a one-to-many mismatch from SL lexicon to TL lexicon. This frames the translation task as a problem of selecting one word out of a limited set of retrieved TL entries, a *listed lexical selection* problem. This classification of the example yields no useful generalizations however: different examples within the same class of "one-to-many" mismatches will require different MT solutions. Consider just one. The English word *know* translates into French *connaître* or *savoir*, a one-to-many mismatch that depends on the semantics of the verb's argument, not a limited selection at all. The one-to-many classification is too general, so that whatever MT solution is developed to handle the *fish/pez-pescado* case, it will have no bearing on this *know/connaître-savoir* case.

A **linguistic** characterization of this example is that it belongs to the narrower class of "unmarked-to-marked/unmarked" mapping mismatches, again from the SL to the TL lexicon. Note that operationally this presents a *lexical description* problem, identifying lexical entries as marked (leaving others unmarked) in order to address the listed lexical selection problem mentioned above. This classification of the example fails in another way, however; it is too narrow. The markedness information may be lexically ambiguous in one language but not in another. Consider *gato,* the Spanish word for both English words *cat* and *tom.* That is, *gato* is ambiguous with respect to markedness, being either unmarked for all cats or marked specifically for males cats.[9] To translate *gato* into English where two words are available, *cat* or *tom,* a MT system requires a *markedness preserving* mapping to ensure where possible that a marked word maps to a marked word and an unmarked word maps to an unmarked word.

Note that while this situation may appear to be the same as with translating *fish/pez-pescado*, it is not. With *fish,* there was no lexically stored ambiguity and so the mapping was "unmarked-to-marked/unmarked." In order to generate the marked case, a MT system must necessarily make use of knowledge available either from the rest of the sentence or from the larger context surrounding the sentence. This search will be driven by the TL lexicon in the generation phase. In the *gato* case, the markedness ambiguity is stored in the SL lexicon and so can be passed along during translation starting

[8] The marked/unmarked distinction is originally from phonology and later was extended into syntax, semantics and learnability theory [Bolinger, 1975].

[9] Markedness ambiguity is not language-specific. The English word *goose,* for example, is also ambiguous. It may be unmarked for sex, as in the sentence *That's a goose, not a chicken* or it may be marked as 'female', as in the sentence *That's a goose, not a gander.*

in the analysis phase. In short, even though partitioning the problem space using linguistic information gets us closer to understanding lexical mismatches, this approach combines the information that we need to distinguish between two distinct classes of MT problems.

Given these examples, the challenge then is to determine which cases to group together into equivalence classes, so that a solution for one example in a partition or problem class will work on all the examples in that class. Our approach is **operational** and seeks to place the different mismatches where they are encountered in our MT system. We do not attempt here to cover the full space of lexical mismatches. Rather we have focused our research on individual partitions from those listed below. In our view of the space of mismatches we have restricted our attention to those that occur at the interface between the interlingua and the target language. We have left aside those that must be resolved by reference to broader contextual knowledge of transactions.[10] Here are a number of ways that the mapping from the interlingua to the target language may present options or problems for the MT system within a language:[11]

- synonymy
  - English *go by car, drive*
  - French *aller à pied, marcher* (go on foot, walk)
- gaps in lexicalization
  - no English word for *go by vehicle*
    but *bus, train, jet* (go by bus, train, jet)
  - no German word for *go*
    but *laufen, fahren* (go on foot, go by vehicle)
- core-overlap
  - English non-causative *fall* and causative *fell*
    but non-causative and causative *break*
  - German non-causative *fallen* and causative *fällen*
    but non-causative and causative *brechen*
- specialization/generalization
  - English *cook, bake, boil*
  - German *kochen, backen, sieden*
- core-overlap and specialization/generalization
  - English *move, take, steal*
  - French *bouger, prendre, voler*

[10]We will not discuss this second type of mismatch here. However we can give a short example that may give the reader a sense of what these mismatches entail. Consider the case of how the Japanese, French and Germans lexicalize the same transaction during a bus ride: *punch the ticket, validate the ticket,* and *invalidate the ticket* [Tsujii and Ananiadou, 1993]. 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 mismatches.*

[11]Some of these examples are taken from related work by [DiMarco *et al.*, 1993]. Their research addresses many of the same questions we are examining.

3.2 Our Focus

While all of the cases listed above have been relevant to our work building semantic ontologies, in the examples that follow we will focus on the lexicalization gaps in particular. 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.[12] These three lexical/ontological mismatches span the possible cases of meaning mismatch [Barnett *et al.*, 1994]. Here we will briefly give an example of each of these for translations from English into German.

In Figure 1 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,* a selection algorithm can opt for a *subsumes* relation to resolve the mismatch and pick this TL word as the head for its translation of *bus.*[13]

A *subsumes* relation is not always available to resolve lexicalization gap mismatches. The non-causative English verb *move* is very general and does not exist as a lexical entry in German. At that level in the ontology, there is no more general non-causative concept to tap for the translation. In this case the selection algorithm must opt for a *subsumes-by* relation by picking a TL word below the SL word 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.

Finally an *overlap* relation occurs in translating the causative English verb *bus* where again we find no corresponding German lexical entry in the ontology. 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 reduction link relation (see lowest dashed arrow in Figure 1) — not a subsumption relation – is needed to capture *go by bus.*

These examples have been introduced here to clarify how lexicalization gaps fall into three lexical/ontological mismatch classes. We will come back to these examples in section 5 when we discuss the algorithm for traversing our semantic ontology in each of these cases.

4 Combining Aspects of Lexical Semantics and KB

Our research in developing a lexical semantics for MT is derived from the LCS formalism of [Jackendoff, 1983].

[12]The closest TL word is *over-general, over-specific,* or *overlapping* in meaning with respect to SL word's meaning.

[13]Suffice it to say here that the selected verb becomes the head of the TL phrase and the information dropped in the move up the ontological hierarchy becomes the restrictive modifier phrase to that head. This is the verb phrase analog to the approach taken by [Sondheimer *et al.*, 1990] with noun phrases.

His work however does not address the **computational** issues associated with representing or composing LCSs.[14] In particular, though Jackendoff writes that thematic relations (i.e. the roles in predicate-argument structure) depend crucially on an enriched ontology, he leaves open to interpretation (i) what that ontology or knowledge base ought to look like and (ii) what would constitute an adequate scheme for encoding the primitives of the LCS representation in the knowledge base. Our approach retains the benefits of the LCS formalism for lexical semantics while augmenting the representation with links into KB concepts. We explain these goals briefly below.

4.1 Benefits of Retaining the LCS formalism

Though several arguments have been made in favor of an LCS-based MT approach, we only cover two here; one benefit relates to IL MT systems in general, and the other bears specifically on capturing mismatches.

One advantage to using an LCS-derived formalism for an interlingua is that its interface to the syntactic component of a MT system has been well developed. The LCS formalism provides a structured representation with predicate-argument forms and the potential for designating operator scoping relations. Because there is a "syntax" to these lexical-semantic representations, it is possible to provide a systematic mapping between the LCS representation and the corresponding syntactic structure. Furthermore, that mapping may be captured in a small set of linking rules, parameterized for cross-linguistic variation. (For an extensive description, see [Dorr, 1993].)

The compositionality of this representation provides us with another benefit that is relevant to lexical selection research: the LCS formalism captures the phenomena of argument incorporation. Since languages differ with respect to what information they incorporate [Talmy, 1985], an IL formalism that retains incorporated information generation will provide a greater variety of lexicalization options at lexical selection time. In order to show how the LCS formalism can do this, we compare how three verbs that have different incorporation properties go through the IL composition and then the lexical selection phase during generation.

Informally incorporation refers to the semantics of one word, such as *lift,* hiding or incorporating the semantics of another word or phrase, such as *up.* Other words, such as *ascend* and *climb,* also contain within them the meaning *up.* This incorporation information is encoded in each verb's RLCS as the same substructure, the RLCS for the word *up.*

At analysis time the *up* substructures in *lift, ascend,* and *climb* are treated differently.[15] The *up* substructure inside of the *lift* RLCS is marked as optional and will absorb an identical structure during composition. When the RLCSs for *lift* and *up* are composed, the resulting composed form is identical to the RLCS for *lift* alone. By contrast, in the RLCS for *ascend,* the *up* substructure is unmarked and inaccessible during analysis.[16] Finally in the RLCS for *climb,* the *up* substructure is marked as a default option and it may be overwritten by non-identical structures during composition, as with *climb down.*

The markings on the RLCSs are removed after the analysis phase is completed, leaving a language-independent CLCS as the interlingual form in the MT system. The portion of the IL forms corresponding to (what was) *lift (up), ascend, climb (up)* all contain the same substructure *up.* In an MT system, the lexical generation step will decide whether or not to lexicalize the equivalent of a TL *up.*[17] The key benefit here is that the LCS formalism has preserved the substructure information, and so does not prematurely foreclose the decision process on lexicalization. The LCS formalism, by leaving the lexicalization decision open into the TL phase, allows for TL-specific pragmatic information to be used and for stylistic choices to be made in the final generation steps[18] — after the lexical options have been identified from the IL form via classification and the semantic ontology has been traversed.

4.2 Encoding the LCS formalism in the KB

We adopt the view that it is possible to consider the status of the LCS primitives from a KR point of view. We take the LCS primitives to be linguistic realizations of KR concepts, i.e., the LCS primitives are grounded in the KB. Part of the motivation for this combined framework is that it allows us to gain precision and accuracy with respect to the KB.

We achieve this LCS/KB melding by assuming (i) the base set of LCS primitives to be a base set of our LOOM-based *lcs-concepts* whose *lcs-roles* (if any) represent the positions in the LCS structures, and (ii) the non-primitive LCSs limited to those appearing as RLCSs in the MT lexicons to be another set of LOOM *lcs-concepts* defined recursively on the LOOM base set. In other words, for every LCS primitive and every non-primitive LCS found in the MT lexicons, there is an lcs-concept.

Recall that RLCSs are configurations of one or more of the LCS primitives composed into a single structure that meets the well-formedness conditions of the LCS "syntax." We may now restate the mapping from the last paragraph in terms of RLCSs: each primitive RLCS maps into one of the KB *base set* lcs-concepts and each

[14] He points this out explicitly in response to criticism from some in the computational linguistic community[Jackendoff, 1992].

[15] If we were to classify these verbs in Figure 1, *lift* would be subsumed by [CAUSE(_ ,[GO-LOC(_ , _)])] and *ascend* and *climb* would be subsumed by [GO-LOC(_ , _)].

[16] We thereby avoid letting the RLCSs for *ascend* and *up* collapse into one RLCS for *ascend:* this lexical *up* does not have the spatial sense of the incorporated *up* substructures.

[17] 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* or *climb up* SL input, there would be a *steigen/aufsteigen* choice (among several 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 synonymy mismatch in section 3.1.

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

non-primitive RLCS maps into one of the KB *derived set* lcs-concepts. Hence any relation that exists among substructures within an RLCS has a corresponding lcs-based relation among lcs-concepts in the KB. The full set of lcs-concepts and their lcs-based relations constitute the semantic ontology of the MT system encoded in the KB.

The next section provides further discussion on classifying RLCS-based lcs-concepts in the semantic ontology and on traversing the semantic ontology during translation runtime. The goal of the remainder of this section is to determine the mapping from relations among RLCSs[19] to relations among LOOM-based lcs-concepts.[20] In particular, we clarify why both subsumption and reduction links are needed in the KB to properly mirror the richness of relations among RLCSs.

Within a non-primitive RLCS, there are, as a result of the recursive definitions in the LCS syntax, two RLCS substructures that comprise C; one which we call the root RLCS (call it R) because it includes the root node, and the other which we call the non-root RLCS (call it N) that does not include the root.

When we map RLCSs into the KB we wish to preserve both the relation of C-to-R and the relation of C-to-N. A look at Figure 1 will help clarify that the relation of C-to-R is encoded in our semantic ontology as what we have been rather casually calling a *subsumption link.* For example, the RLCS for *move* in the left-hand column stands in a C-to-R relation with the RLCS for *cause.* We capture this in the KB by saying that *move* is subsumed by *cause.* The relation of C-to-N is encoded as a *reduction link.* For example, the RLCS for the causative *move* stands in a C-to-N relation with the RLCS for the non-causative *move;* the non-causative *move* is reached via a reduction link from the causative *move.* As discussed in the next section, the reduction links serve to partition the lcs-concept being translated into subcomponents where each has its meaning preserved.

5 LOOM Implementation

There are two stages to our solution to the lexical mismatch problem. The first involves the classification of words according to their conceptual description prior to translation runtime. The second involves the selection of words during translation runtime based on the conceptual representation that comprises the IL representation. Each of these will be described, in turn.

5.1 Before Lexical Match Time

Our approach requires the semantic lexicographer to create conceptual descriptions (RLCS's) embodying the meaning of lexical entries.[21] As discussed earlier, these representations are stored in their own language-specific lexicons. When the RLCS definitions are loaded, LOOM automatically infers the ontological relationships between their concepts, thus creating a hierarchy[22] into which each language's lexical entries point. Figure 2 illustrates the LOOM specification of the ontology presented earlier in Figure 1.

Note that lcs-roles in Figure 2 correspond to thematic positions (underscores) in Figure 1. Concepts and their language-specific instances are prefixed by "lcs-"; the instances are suffixed by "-i". The RLCSs of the language-specific lexicons in Figure 1 and their LOOM-encoded versions of the semantic ontology in Figure 2 have been simplified here for ease of presentation.

5.2 Lexical Selection and Mismatch Handling

This section contains a functional description of the lexical selection process and some associated mismatch handling that the TL classification hierarchy buys us. For the inner workings of the algorithm, see section 5.3 below. We discuss the three mismatch cases mentioned earlier in section 3.2. (In this context "traversing" will refer to the search procedure at translation runtime that is invoked following the initial mismatch diagnosis returned by the term classifier.)

For the purpose of this discussion, it will suffice to focus on the translation of single words rather than an entire sentence. This process is intended to operate recursively, i.e., once the TL verb is selected the same lexical selection procedure is applied to the arguments of that word.

5.2.1 Resolution of Subsumption Cases

¿From a high level view, the treatment of both the *subsumes* and the *subsumed-by* cases are similar. Both cases produce lists of word concepts: the first list simply contains the ontological parents of the CLCS we look up in the TL hierarchy, while the second list contains the ontological children. The subsumed-by case corresponds to a situation where the closest matching TL word is subsumed by the SL word concept. The subsumes case corresponds to a situation where the closest matching TL word subsumes the SL word concept.[23]

Consider first the non-causative English verb *bus* in figure 1, as used in the sentence *they bussed into the city.* As mentioned earlier, there does not exist an equivalent lexical entry among the verbs in German (which we have annotated in Figure 2 with a '***'). The best translation, *Sie fuhren mit dem Bus in die Stadt,* uses the combination of the verb *fahren* (go by vehicle) and the prepositional phrase *mit dem Bus* (by bus) to convey a comparable meaning. In our interlingual representations, this mapping to the German verb *fahren* from

[19] Here we use RLCS to refer to the Jackendoffian encoding of lexical entries as seen in simplified form in Figure 1.

[20] Here we use lcs-concepts to refer to the LOOM-based encoding of RLCSs.

[21] We currently have a database containing 250 RLCS templates for English typed in by hand on the basis of work by [Levin, 1993]. We expect to automatically acquire 3200 English verbs using these templates. (Preliminary results are in [Dorr *et al.*, 1994a].) Once this acquisition process is complete, we are planning to scale up the current experiment so that we can classify this larger set in terms of the LOOM representation.

[22] We use the term *hierarchy*, in a loose sense which does not exclude multiple inheritance.

[23] In LOOM, we use the commands direct-superconcepts and direct-subconcepts for reaching immediate ancestors and descendants, respectively, in combination with the command superconcepts to test for other adequate TL word concepts.

the English verb *bus* corresponds to the subsumes case: the TL meaning [GO-LOC(_ , _), (BY VEHICLE)] subsumes the SL meaning [GO-LOC(_ , _), (BY BUS)]. In the semantic ontology, this corresponds to a straightforward traversal step upward along a subsumption link. Note that although this link is within the ontology constructed with LCS concepts), it is grounded in the general, i.e. non-LCS ontology, part of the KB where the concept vehicle subsumes the concept bus.

Next consider the non-causative English verb *move,* as used in the sentences (i) *the tree moved* and (ii) *the car moved through the tunnel.*[24] This very general sense of *move,* which we encode as GO-LOC in the RLCS to indicate a change of spatial location, does not translate directly into German. At this level in the ontology, there is no more general non-causative concept to tap for the translation. Consequently the best translations for (i) and (ii) respectively, *der Baum bewegte sich* and *das Auto fuhr durch den Tunnel,* make use of less general verbs that, by definition, capture a narrower meaning. With *sich bewegen* the nature of the movement is constrained, and with *fahren* the logical subject of the movement is constrained.[25] These mappings to *sich bewegen/fahren* from *move* correspond to a traversal step down a subsumption link from the SL word concept in the ontology.[26]

5.2.2 Resolution of Overlapping Cases

By far, the most interesting case is that of the *overlapping* case. Whereas the subsumes and subsumed-by cases are covered by the Directed Subsumption Links, the overlapping cases require the use of Crossover Reduction Links, which constitute the main contribution of our representation scheme.

This case is more complex that the others in that the TL and SL meanings may overlap in several ways, some of which cannot be classified strictly in terms of subsumption. Several translations are possible, reflecting these many overlapping combinations. As mentioned earlier, the English causative verb *bus,* as used in the sentence *the parents bussed the children to school,* has no German lexical equivalent. One German translation of this sentence that conveys an official or formal style is *die Eltern veranlassten die Kinder, mit dem Bus zur Schule zu fahren* (the parents caused the children to go to school by bus). In our interlingual representations, the meanings for the German words *veranlassen* [CAUSE(_ , _)] and *fahren* [GO(_ , _), (BY VEHICLE)] each overlap the English meaning for *bus* [CAUSE(_ ,[GO-LOC(_ , _),(BY BUS)])]. (See lower left corner of figure 1.) While the causative *bus* is subsumed by *veranlassen,* there is a reduction link relation in the traversal from *bus* to *fahren.*

Figure 2: LOOM Version of RLCS Classification

[24] We identify these usages of the word *move* as non-causative in the broadest sense that the cause or causer of the movement is unknown.

[25] In terms of figure 2, *sich bewegen,* were it included, would be a child of the lcs-go concept.

[26] At this stage in generation, we are concerned with extracting lexicalization options, building a set of one or more RLCSs in the TL. At the subsequent step of TL RLCS composition, the constraints defined within each LOOM-version of the TL RLCS will be checked against each other.

Note that there often arise several possible translations in the overlapping cases precisely because there are several ways to lexicalize this overlapping concepts. The search procedure that traverses the ontology explores several paths trying to find the set of matches from the German lexicon that covers the IL form. In general, reduction link traversals are preferred over subsumption link traversals since reduction links preserve exact substructures whereas subsumption links require inferencing.

5.3 Inner Workings of the Lexical Selection Algorithm

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.

The details are given here:

- **CLCS Lookup:** Search the ontology on the basis of the IL representation (i.e., the CLCS of [Dorr, 1993]), in order to determine its hierarchical position[27]
- **RLCS Extraction:** Examine instances of TL words that occur at the current hierarchical position.
- **Merge Detection:** In the cases where there exists an exact match between a TL RLCS and the IL structure, the associated lexical item is returned.[28]
- **Traversal of Semantic Ontology:** In the cases where there is no exact match between a TL RLCS and the IL structure, a search procedure is invoked to traverse the LCS concept ontology, following subsumption and reduction links. Along each search path, after traversing each link, evaluate returned TL RLCSs for coverage of IL form.
  - **Overlapping Case:** The reduction links are followed in order to find a set of possible substructures that provide full coverage for the concept. Return a list of associated reduction RLCSs.
  - **Familial Information Case:** Subsumption links are followed, upward by default, downward if this is not possible because traversal is at top of LCS subsumption hierarchy. Return the list of ontological parents (if upward traversal) or list of ontological children (if downward traversal).

Note that overlap reduction has a higher standing than subsumption. This is because our goal is to get the fullest possible coverage of the IL concept. With reduction, we only partition the IL form in terms of its coverage by a TL RLCS. With subsumption, we may lose information in the inferencing that would have led to a more accurate lexicalization.

[27] When the CLCS corresponds to a SL word's RLCS, the search will find a match in the ontology corresponding to the concept where the SL was classified before translation runtime. What is at issue in the next step is whether, for this SL concept just looked up, the TL will have a word defined as an instance off of that same concept.

[28] This was the case for the example given in section 2 translating the German *bewegen* into English *move*.

6 Limitations and Conclusions

Future work will include the detection of non-subsumption/overlap cases, i.e., where differences are not resolved solely by means of appropriate word selection but rather by means of constraints on relations defined by the roles encoded in word definitions. For example, one could envision a language where the concept *steal* could only be realized as some combination of the concepts *take* and *unauthorizedly*. Suppose these were classified in the ontology, but that there were no word for the concept *unauthorizedly*. Constraints on the relations between roles would allow us to translate *unauthorizedly* as something akin to *without permission*, perhaps as an instantiation of the concept *movement of an item by a person that is not a member of the item's list of authorized custodians*.

An additional area for future investigation is the use of knowledge within a KB to assist in filtering. For example, if we were translating the sentence *Jane bought a fish from the pet shop* into Spanish, we would need to have knowledge about pet shops, i.e., that a pet shop sells pets and supplies. With deeper knowledge, the word *pescado* would be filtered out because the associated concept (a fish that has been caught to be eaten) would not unify with purchase activities entailed by a pet shop.

We have presented a concept-based approach to the problem of lexical selection. One important result is the ability to deal precisely with the case where the target language lacks a word that directly matches the IL fragment to be generated in the target language. Our scheme for handling lexical mismatches is linguistically motivated in that it makes use of a decompositional representation that has proven useful for structural realization of the TL sentence [Dorr, 1993]. The novelty of our approach is that it addresses the problem of mismatches through access to the KR while retaining enough of the structure of the IL to resolve surface-level distinctions. A main contribution of our representation scheme is the use of Crossover Reduction Links for handling cases where two concepts overlap, but neither subsumes the other. We have demonstrated the utility of our approach for mismatched verbs in English and German.

References

[Barnett *et al.*, 1994] J. Barnett, I. Mani, and E. Rich. 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 Publishers, 1994.

[Beaven, 1992] J. Beaven. Shake and bake machine translation. In *Proceedings of Fourteenth International Conference on Computational Linguistics*, pages 603–609, Nantes, France, 1992.

[Bolinger, 1975] Dwight Bolinger. *Aspects of Language.* Harcourt Brace Jovanovich, New York, NY, 1975.

[Brachman and Schmolze, 1985] R.J. Brachman and J. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216, 1985.

[DiMarco *et al.*, 1993] C. DiMarco, G. Hirst, and M. Stede. The semantic and stylistic differentiation of synonyms and near-synonyms. Technical Report AAAI-93 Spring Symposium on Building Lexicons for Machine Translation, Stanford University, Stanford, CA, 1993.

[Dorr and Voss, 1993] B.J. Dorr and C. Voss. Machine translation of spatial expressions: Defining the relation between an interlingua and a knowledge representation system. In *Proceedings of AAAI–93*, 1993.

[Dorr *et al.*, 1994a] B.J. Dorr, J. Garman, and A. Weinberg. From subcategorization frames to thematic roles: Building lexical entries for interlingual MT. In *Association for MT in the Americas Conference on Partnerships in Translation Technology*, submitted, Columbia, Maryland, 1994.

[Dorr *et al.*, 1994b] B.J. Dorr, D. Lin, J. Lee, and S. Suh. A paradigm for non-head-driven parsing: Parameterized message-passing. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

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

[Dorr, 1994] B.J. Dorr. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4), 1994.

[Iwanska, 1993] L. Iwanska. 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, 1993.

[Jackendoff, 1983] R. Jackendoff. *Semantics and Cognition*. MIT Press, Cambridge, MA, 1983.

[Jackendoff, 1992] R. Jackendoff. What is *Semantic Structures* about? *Computational Linguistics*,18(2), 240–242, 1992.

[Kameyama *et al.*, 1991] M. Kameyama, R. Ochitani, S. Peters, and H. Sirai. 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, 1991.

[Kinoshita *et al.*, 1992] S. Kinoshita, J. Phillips, and J. Tsujii. Interaction between structural changes in machine translation. In *Proceedings of Fourteenth International Conference on Computational Linguistics*, pages 679–685, Nantes, France, 1992.

[Levin and Nirenburg, 1993] Lori Levin and Sergei Nirenburg. 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, 1993.

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

[Lindop and Tsujii, 1991] J. Lindop and J. Tsujii. Complex transfer in MT: A survey of examples. Technical Report CCL/UMIST Report 91/5, Center for Computational Linguistics, UMIST, Manchester, UK, 1991.

[MacGregor, 1991] R. MacGregor. The evolving technology of classification-based knowledge representation systems. In J. Sowa, editor, *Principles of Semantic Networks*, pages 385–400. Morgan Kaufmann, San Mateo, CA, 1991.

[Melby, 1986] A.K. Melby. Lexical transfer: Missing element in linguistic theories. In *Proceedings of Eleventh International Conference on Computational Linguistics*, Bonn, Germany, 1986.

[Nirenburg and Nirenburg, 1988] S. Nirenburg and I. Nirenburg. A framework for lexical selection in natural language generation. In *Proceedings of Twelveth International Conference on Computational Linguistics*, pages 471–475, Budapest, Hungary, 1988.

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

[Quantz and Schmitz, 1994] J.J. Quantz and B. Schmitz. Knowledge-based disambiguation for machine translation. *International Journal of Minds and Machines, Special Issue on Knowledge Representation for Natural Language*, 4:39–57, 1994.

[Schubert, 1991] L. Schubert. Semantic nets are in the eye of the beholder. In J. Sowa, editor, *Principles of Semantic Networks*, pages 95–107. Morgan Kaufmann, San Mateo, CA, 1991.

[Shapiro, 1993] A. Shapiro. 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, 1993.

[Sondheimer *et al.*, 1990] N. Sondheimer, S. Cumming, and R. Albano. How to realize a concept: Lexical selection and the conceptual network in text generation. *Machine Translation*, 5(1):57–78, 1990.

[Sowa, 1991] J. Sowa. Toward the expressive power of natural language. In J. Sowa, editor, *Principles of Semantic Networks*, pages 157–189. Morgan Kaufmann, San Mateo, CA, 1991.

[Stede, 1993] M. Stede. Lexical options in multilingual generation from a knowledge base. Technical Report Manuscript, University of Toronto, Toronto, Canada, 1993.

[Talmy, 1985] L. Talmy. Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen, editor, *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon*,pages 57–149. Cambridge University Press, Cambridge, 1985.

[Tsujii and Ananiadou, 1993] J. Tsujii and S. Ananiadou. Knowledge-based processing in MT. Manuscript, U.S. - Japan Machine-Aided Translation Workshop, Washington DC, November, 1993.

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

[Woods and Brachman, 1978] W.A. Woods and R.J. Brachman. Research in natural language understanding. Technical Report Quarterly Technical Report Progress Report No. 1, Bolt, Beranek, and Newman, Cambridge, MA, 1978.