Constraining Lexical Selection Across Languages Using Tree Adjoining Grammars

Dania Egedi Chunghye Han Fei Xia
Martha Palmer and Joseph Rosenzweig*
Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia PA 19104-6228
{egedi, chunghye, fxia, mpalmer, josephr}@linc.cis.upenn.edu

1 Introduction

One of the primary tasks for Machine Translation (MT) is lexical selection — selecting the target lexical item that most closely matches the source lexical item being translated. For transfer-based approaches Geta (Vauquois and Boitet, 1985), each separate lexeme in the source language must be paired with a corresponding lexeme in the target language in a set of bilingual dictionaries. An alternative [1] is the interlingua approach, such as Princitran (Dorr, 1993) or Translator (Nirenburg *et al.*, 1992), in that the source verb is mapped to a canonical semantic representation which is shared by all target languages. The elements of the semantic representation select the lexical realization in each target language.

There are several components of lexical selection. Most MT applications are small or relatively domain specific, so an important aspect of lexical selection is generally overlooked — distinguishing between lexical items that are closely related conceptually. There can be many shades of distinction between the meaning of a lexical item in one language and its counterpart in another language (Pye, 1993). These distinctions are sometimes critical to selecting the correct lexical item in the target language. The question then arises, in both transfer and interlingua based systems, of how and where to capture these distinctions. While some MT systems have relegated this task to a 'world knowl-

*We would like to thank Aravind Joshi, Sadao Kurohashi, and Zhibiao Wu for their helpful input. This work was supported by the Center for Command, Control, and Communications Systems (C3) (Mr. George Yaeger) under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle (Delivery Order 1326, Contract No. DAAL03-91-C-0034) and NSF Science and Technology Center Grant SBR 8920230.

[1] These are the two ends of the spectrum, and many systems now take a hybrid approach. Since the purpose of this paper is to highlight a area of MT usually ignored, and to propose a non-theory specific solution, we will not give an overview of all types of MT systems. We do limit our initial comments to non-statistical MT methods, as we do not believe that our method would be useful to purely statistical systems.

edge' or 'pragmatics' module (Carbonell *et al.*, 1981, Sun, 1992), we are interested in seeing how much we can accomplish using a combination of syntax and lexical semantics. In this paper, we outline a proposal to capture these distinctions based on separate ontologies for each individual language. Our method is applicable to both transfer and interlingua based approaches, and provides a more elegant solution than exhaustive enumeration and a more local solution than reliance on 'world knowledge' modules. This method has been implemented in FB-LTAGs (Joshi *et al.*, 1975, Schabes, 1990, Vijay-Shanker and Joshi, 1991), whose feature-based, lexicalized approach provides an advantageous environment for modelling the more specific and language dependent syntactic and semantic distinctions necessary to further filter the choice of the lexical item.

2 Defining the Problem

The essence of the problem that we are trying to solve involves lexical constraints that are critical for one language but non-existent or completely different in another. A classic example of this is the translation of *wear* into Japanese.

(1) karewa boushiwo kaburu.
he hat wear
He wears a hat.

(2) karewa kutsushitawo haku.
he socks wear
He wears a pair of socks.

Sentences (1) and (2) highlight a situation in which one language (Japanese) distinguishes several senses of a concept /WEAR/ that has only one sense in another language (English). In Japanese, *kaburu* selects for items worn on the head, such as *hats*, while *haku* selects for items such as *socks*. English *wear* does not make this lexical distinction.

A similar situation is illustrated by the following Korean examples from a military domain context.

(3) choykunuy yocheng-ul swusinha-yss-ta
current request-ACC receive-past-IND
I received the current request

(4) kongkupmwul-ul pat-ass-ta.
commander-report-ACC receive-past-IND
I received the supplies.

Sentences (3) and (4) highlight a situation in which one language (English) has two senses for the same lexical item, *receive*, whereas the other language, Korean, has two distinct lexical items corresponding to these

same senses.[2] In Korean, the first sense of *receive* is *swusinhayssta* and it selects for a theme argument which denotes some information such as *request* or *command* that is transmitted via a communicative device such as a radio transmitter or a telephone. The second sense of *receive* is represented by *patassta*. The sense of *patassta* is more like that of English *receive* in that it allows a wider range of theme arguments. That is, the theme argument of *patassta* can denote physical objects such as *supplies* as well as *information* such as *report*. Hence, in translating *We received the supplies* into Korean, the corresponding verb for English *receive* must be *patassta*. However, in translating *We received your request*, the corresponding verb for English *receive* should be *swusinhayssta*. Hence, selectional restrictions of verbs must be specified in such a way as to block wrong translations.

Translating the English lexical item *break* into Chinese presents an even more complex set of issues, since Chinese offers literally dozens of expressions for describing breaking events, each one of which is more specific than *break*. This causes difficulties for a large-scale transfer-based system such as TRANSTAR, a commercial broad coverage English/Chinese MT system developed in Beijing. When this system was applied to sentences from the Brown corpus that contain *break*, an accuracy rate of less than 20% was achieved, even after ruling out idiomatic uses and problems with parsing (Palmer and Wu, 1995). The primary reason is that in English *break* can be thought of as a very general verb indicating an entire set of breaking events which can be distinguished by the resulting state of the object being broken. *Shatter*, *snap*, *split*, etc. can all be seen as more specialized versions of the general breaking event. Chinese has no equivalent verb for indicating this class of breaking events, and each usage of *break* has to be mapped onto a more specialized lexical item. Even the English specializations of a breaking event do not cover all of the different ways in which Chinese can semantically distinguish between breaking events. The end result is that lexical selection from English to Chinese is often predicated on the existence of semantic features that are completely irrelevant to English.

This is not a problem that is unique to translating between English and Asian languages. In looking for cross-linguistic semantic universals for *break* and other semantically similar verbs, Pye found that there were as many different semantic classification schemes as there were languages being investigated (Pye, 1993). The solution to this problem is elusive enough when considering two particular languages. It must be recognized that a typical transfer-based approach requires a direct mapping from each distinct verb sense to its corresponding lexical item in the target language. In order to achieve this type of mapping for the previously mentioned *break*

[2] Because Korean is a pro-drop language, the target translations do not contain a lexical item corresponding to the English subject pronoun *I*. Though this poses additional complications to the translation process, it does not bear directly on the problem being addressed here.

examples, the English lexical item *break* would have to be subdivided artificially into several distinct lexical items, i.e., *break1, break2, etc.*, using the semantic features that are relevant to Chinese so that each distinct Chinese expression would have a corresponding English expression. In other words, *break1* would map to *da sui* and would have brittle as a semantic feature on the verb object, *break2* would map to *da duan*, and would have a semantic feature line-segment-shape on the verb object, and so on. The semantic features that are relevant to Chinese have to be incorporated into the English lexicon, and vice versa, to establish the accurate correspondences. Each lexicon must therefore specify all of the semantic features relevant to both languages. The interlingua approach has a similar difficulty, since it must define an interlingua that can capture all of the semantic features for both languages. When one begins to consider the problem from the perspective of several languages, this technique quickly becomes impractical. The direct mapping approach becomes cumbersome, unwieldy, and extremely tedious to build, since it means reanalyzing the semantic features of each language according to every language that it is being paired with. For the interlingua, a vast, language universal ontology must be built that incorporates every semantic feature for every language in an organized fashion. That means that not only do correspondences have to be found between individual lexical items, but also between the classification schemes by which each language structures its concepts. While there has been a lot of promising recent work on the problem of verb classification, it is not clear that it supports the notion of a readily accessible language universal ontology. For instance, Levin (Levin, 1993) has shown that there is a correspondence between lexical-semantic verb classes and syntactic structures for English and there has been speculation that these verb classes should extend to other languages since they are based on cross-linguistic semantic concepts. Mitamura, however, has determined a classification for Japanese verbs that shows very little correspondence to Levin's classes (Mitamura, 1989). The EDR project, an enormous effort (over 200,000 words) to build an English-Japanese bilingual dictionary based on a joint conceptual classification, has found a conceptual overlap between the two languages of only about 10% (Yasuhara, 1993). Another large ongoing effort in France has also been looking at generalizations about verb classes in French that can be made based on allowable syntactic transformations. This work is currently being extended to several other languages, but each language is being done independently, from the ground up, with very little sharing of classification schemes (Leclerc, 1990). None of this rules out the possibility of semantic universals, or large areas of conceptual overlap between languages, but it does highlight the extreme individuality of each language, and the overwhelming task that lies in front of anyone trying to merge language-specific conceptualizations.

3 Proposed Model

We believe that the most practical approach is to assume that each language will require its own conceptual ontology with a distinct set of semantic features. Many of the concepts in the lexical semantic ontologies may be shared among languages, but languages may choose to structure the concepts differently. With this in mind, we suggest an approach to translation that does not always attempt to directly map a specific verb sense in the source language to another specific sense in the target language. Rather, it begins with a more coarse-grained lexical translation process, which merely attempts to focus on a particular set of translation candidates in the source language. These candidates will be further narrowed down by a language-specfic lexical selection process which examines the semantic features associated with the instantiated verb arguments and determines the best fit. Therefore, in many cases, the detailed merging of language-specfic semantic features associated with the source sense into the semantic features of the target sense can simply be avoided. Rather than one-to-one mappings between lexical items, the dictionary would map between sets of lexical items. We see this as a hybrid approach that combines some of the strengths of both interlingua-based systems and transfer-based systems. In an interlingua system, the goal is to capture semantic similarities by associating several lexical items with the same primitive concept. This is equivalent to assigning these lexical items to the same class. We also see semantic classification as a critical component of our lexical organization, but we do not expect the classes to be uniquely defined by a small set of language universal semantic primitives. Neither do we expect the class membership to be a substitute for the specific representation of the individual lexical item. We will be retaining the complete semantic representations with selectional restrictions for the individual lexical items. We see the group "class" or "concept" as an enhancement of this representation, which will play a crucial role in selecting potential translations. Any single lexical item might belong to several different classes, which may only have two or three other members. Semantic classifications that play pivotal roles when translating between languages A and B may be of only minor importance to translations between languages C and D. Determining these subtle shifts in relevance among alternative classification schemes is the most difficult part of our task, and will require access to vast amounts of data, preferably alligned bilingual corpora.

From the perspective of the transfer approach, the biggest difference is a broader, less fine-grained, mapping between the source language and the target language. Instead of always mapping directly from an individual lexical item to another individual lexical item, the mappings can also be between sets of lexical items. Since the language-specific semantic features are kept local to each language, they do not have to be incorporated into

the feature set of every language. For instance, English *receive* maps to a set of Korean verbs such as *swusinhata* and *patassta*. Correspondingly, *patassta* would map to a set of English verbs such as *get, receive, obtain*. The final selection of the actual lexical item will be made in the target language based on the semantic features associated with the prospective arguments, and eventually on pragmatic factors as well. One of the advantages of our approach is that the same self-contained, language-specific representation that is normally used for any form of analysis or generation becomes very applicable to machine translation (Palmer *et al.*, 1993). More importantly, it is not necessary that the languages being translated have the same underlying verb classes, since the semantic structure is local to each language. However, we cannot entirely avoid the issue of finding the conceptual links between language-specific classification schemes. We are still left with the problem, given the different classification schemes, of associating appropriate classes of lexical items in a target language with the most closely corresponding classes in the source language. Since we have just argued that there will never be exactly corresponding classes in any two languages, this is clearly still a difficult issue. However, we do not have to try to force the different classification schemes into a single interlingua. As discussed above, it might be that the most useful method for taking advantage of our approach would be in a hybrid system that uses a direct transfer method whenever possible, and a more general, classification-correspondence approach in other circumstances, as described in Palmer and Rosenzweig (Palmer and Rosenzweig, 1996).

4 Implementation

Our implementation of this model uses a variant of the Synchronous TAGs formalism, a Lexicalized TAG suitable for machine translation (Shieber and Schabes, 1990, Abeillé *et al.*, 1990), which has been augmented to handle feature-based unification. This particular formalism has a number of advantages for our approach. First, it is lexicalized, which makes it easier to specify the lexically specific semantic information in a syntactic context. This is important in languages such as English where the semantics can have syntactic consequences (Levin, 1993). Second, it is feature-based, which provides a convenient notation and mechanism in which to specify the selectional restrictions. Third, the extended domain of locality provided by the tree structures allows lexical items to easily place constraints on other lexical items in the same frame.

Syncronous TAG adopts the strategy of matching the source FB-LTAG derivation of the source sentence to a target FB-LTAG derivation by looking in a transfer lexicon. The transfer lexicon consists of pairs of trees from the source grammar and target grammar. Within a pair of trees, nodes may be linked. The translation process is outlined in (Abeillé *et al.*, 1990).

First, the source sentence is parsed using the source grammar. Second, the source derivation is transferred to a target derivation by mapping each elementary tree in the source derived tree to a tree in the target derived tree. This is done by looking in the transfer lexicon. Finally, the target sentence is generated from the target derived tree. As an example, we provide a fragment of the transfer lexicon between English and Korean, and show how the English sentence *John likes Mary* is translated into the corresponding Korean sentence.

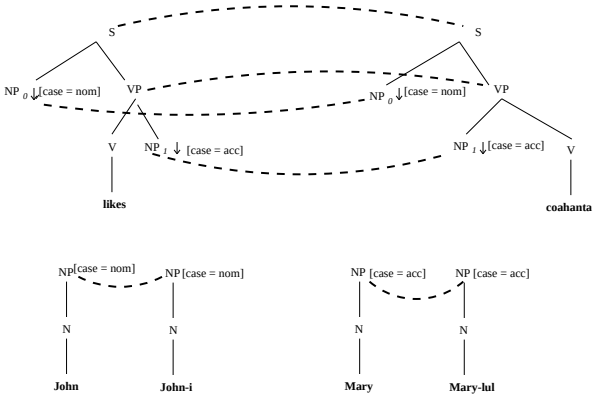


FIGURE 1 Lexicalized Synchronous trees for *like*, *John*, *Mary*

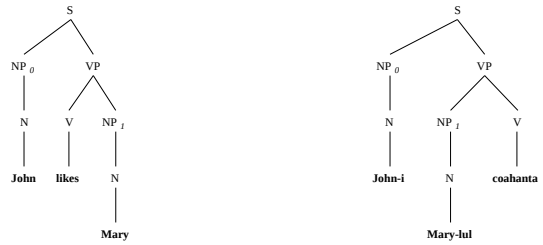


FIGURE 2 *John likes Mary* translated into *John-i Mary-lul coahanta*

Figure 1 shows the links between elementary trees for English and Korean. After the English sentence *John likes Mary* is parsed under the English grammar, the derived tree is transferred to a target derived tree by mapping each elementary tree in the source derived tree to a tree in the target derived tree. Finally, the target sentence *John-i Mary-lul coahanta*[3] is generated from the target derived tree. Figure 2 shows the source and target derived trees.

Although Synchronous TAGs can also be used with an interlingua

[3] *-i* is a nominative case marker and *-lul* is an accusative case marker.

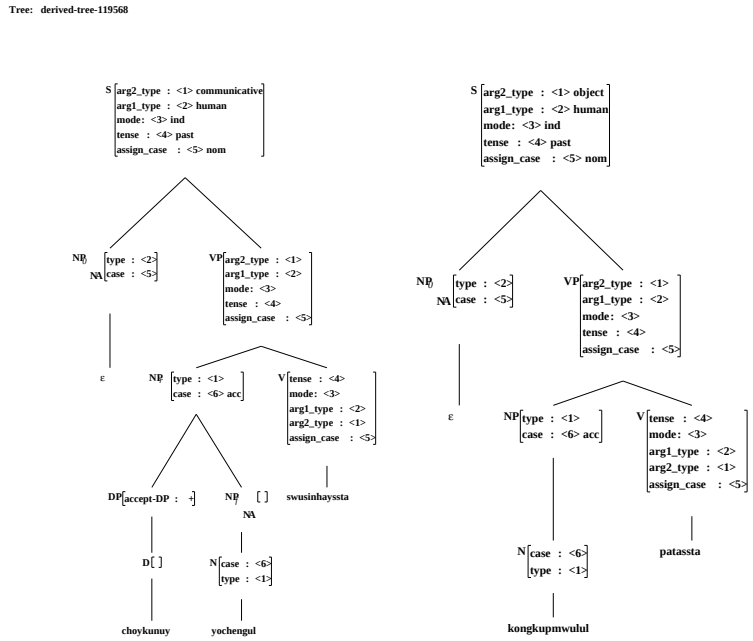
(Sun, 1992), (Dorr and M.Palmer, 1995), after exploring both interlingua approaches and transfer approaches, we have chosen a transfer-based approach for Korean and English. This is partly because of the difficulties we mentioned earlier with respect to handcrafting a truly universal interlingua, but there are other reasons as well. We discovered when translating many of the Lexical Conceptual Structure (LCS) interlingua representations into TAGs that the interlingua would duplicate almost exactly the structure of either the source language or the target language. So whereas the mapping from one side would resemble very closely what would be involved in a transfer mapping, the mapping from the other side would be quite trivial. We soon began to feel that the interlingua was giving us an unecessary extra step, especially when it also became clear the the actual lexical item for the source language (always English) had to appear explicitly somewhere in the LCS representation. The LCS seemed to be simply associating three things together - the syntactic structure of either the source language or the target language, a set of semantic components, and a particular lexical item. We felt that the same associations could be made more efficiently by using the target language syntactic structure anchored by the lexical item and annotating it with necessary semantic components.

We will work through several examples to show how Synchronous FB-LTAGs handle this method. Semantic constraints are specified in the usual method for each language. The semantic characteristics of a lexical item (or each sense of a lexical item) are instantiated as features in the syntactic lexicon. A lexical item may also specify constraints on semantic features of other lexical items available in its syntactic frame (i.e., local to its tree). At parse time, of course, the features and feature constaints must unify. Since this is done independently for each language, there is no need to access a universal ontology in order to make the lexical selection. The language-specific selectional restrictions will ensure the suitability of the final verb argument structure.

This approach is being successfully applied to a domain of military messages, with English and Korean as our two languages, (Egedi *et al.*, 1994). This effort is being funded by CECOM at Ft. Monmouth. In order to obtain data for this application, we have visited the 75th Division Training Exercise in Houston that involved Fort Lewis in a Corps Battle Simulation where in addition to becoming familiar with the Battle Simulation environment, we were able to collect hundreds of messages, both computer generated and hand-written. In this domain, short telegraphic messages are sent to military units with requests for information and supplies, and corresponding answers are sent as replies. The goal is automatic, on-line translation of these messages. We are finding that the approach to lexical selection outlined in this paper is adequate for the lexical choice issues that arise in this domain.

The trees in Figure 3 show the derived trees for the sentences in (3) and

(4) above. The trees in Figure 4 show the NP trees for the argument NPs used in the sentences:



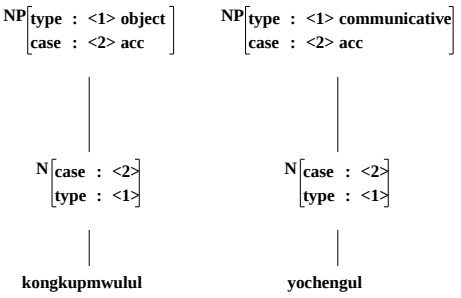
swusinhata tree *patassta* tree

FIGURE 3 Two trees corresponding to English sentences with *receive*

The verb *swusinhayssta* requires a theme argument which denotes something that is transmittable via a communicative device such as a telephone or a radio. This is indicated by the feature *communicative*. The noun *yocheng* denotes something which can be transmitted via a communicative device and so it has the feature *communicative* on the noun tree. The features of the verb and the argument NP are compatible and so the Korean parser accepts the input and generates the correct derived tree.

The verb *patassta* requires a theme argument which denotes a physical object or some information. This disjunctive constraint can be implemented in the TAG formalism as the disjunctive feature-value *information/object*, which indicates that the verb can take both types of arguments. The noun *kongkupmwul* denotes a physical object. Hence, we implement the feature-value *object* on its noun tree. The features of the verb and the argument NP are compatible and so the parser accepts the input and generates the correct derived tree.

In translating from English to Korean, the semantic features implemented for the Korean verbs and nouns ensure that the correct target



supplies tree *request* tree

FIGURE 4 NP trees for Korean *supplies* and *request*

sentence is generated. In the case of the English sentence *I received the current request*, the English verb *receive* correctly maps onto the Korean verb *swusinhayssta*. Also, in translating the English sentence *I received the supplies* into Korean, the English verb *receive* correctly maps onto the Korean verb *patassta*. Since the semantic type of the object of *receive* is not restricted in this way in English, there is no need to implement these semantic features in the English lexicon.[4] The trees for *request* and *supplies* (shown in Figure 5) in English therefore are not marked for their objecthood nor for their ability to be transmitted over a communicative device.

The English grammar possesses only those features which are required within the English grammar itself; the presence of features in the Korean grammar (or grammars for other languages) does not mandate their presence on the English side. Conversely, features relevant for English may not show up in the Korean grammar.

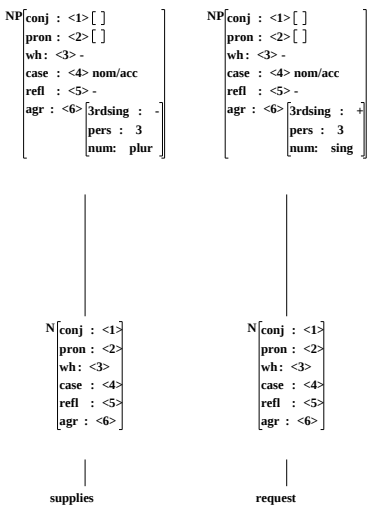
It is important to note that the semantic features of a noun are not always context independent. For example, the noun *poko* (*report*) denotes the information conveyed by an act of reporting. Hence it is compatible with the verb *patassta* (5) — but only in a context where the information was conveyed via face-to-face interaction. If instead the information was conveyed via a communicative device, then *swusinhayssta* is the appropriate choice (6).[5]

(5) poko-lul pat-ass-ta
report-ACC receive-past-IND
I received the report.

(6) poko-lul swusinha-yss-ta

[4]This does not mean that the semantic type features *object* or *information* might not be relevant for other reasons elsewhere in the English grammar.

[5]If the Korean noun *poko* is followed by the morpheme *-se*, then it refers to a physical document that contains the information. Hence, it refers to a physical object and can not occur with the Korean verb *swusinhayssta*. It can only occur with *patassta*.



supplies tree *request* tree

FIGURE 5 NP trees for English *supplies* and *request*

report-ACC receive-past-IND
I received the report.

The current implementation cannot incorporate the kind of discourse context that is crucial in determining the correct translation for the English *receive* for examples such as this one. It is currently still a small, preliminary Korean grammar, with 11 tree families for 30 verbs, 150 nouns, and some function words. There are around 25 entries in the transfer lexicon.

5 More complex examples

The first section, briefly, described the difficulties in automatically translating *break* from English to Chinese, primarily because the Chinese expressions need to be more semantically precise. Not only do they make more explicit the resulting state of the *broken* object, whether it is in small pieces, or pieces shaped like line segments, etc., but they make explicit the action that resulted in the *breaking* event, such as hitting or shouldering.

If the basic meaning for English *break* is *X (the agent) exerts a force on Y (the patient) and causes Y to separate into pieces*, its Chinese translation consists of two parts: the first part (action expression) describes how the agent exerts a force on the patient, the second part (result expression) gives the consequence of the action. The Chinese expression produced is actually a productive compound construction that makes explicit first the action and then the result. Since English *break* is a very general word, it does not make explicit the action that results in the breaking event in the way that Chinese does.

For the result expression, there are dozens of Chinese words which describe the state *into pieces*. The attributes of the patient will decide which result is most likely to occur. For example, a stick can be broken into line segments, a vase can be broken into small pieces. The correct lexical choice for the result part can often be made based on inherent characteristics of the object involved (Table 1).

RESULT OF *BREAK*	RESULT EXPRESSION	ATTRIBUTE OF PATIENT	EXAMPLES
into tiny pieces	*sui*	brittle, physical	window, vase
into large pieces	*po*	solid, physical	window, door
into line segments	*duan*	long, slender, physical	branch, stick

TABLE 1 Correspondences between the patients of actions and the result expressions

Determining the action part is more difficult and often depends on contextual factors that may not be available to a machine translation system. Nevertheless, the action expression in a Chinese sentence is also often heavily dependent on context-independent lexical-semantic information about the types of nouns in the sentence. For example, if an instrumental adjunct phrase is present, the type of the noun in that phrase constrains the action expression that will be used (Table 2). One approach is to avoid committing to a specific action, and use the most general phrase available, as exemplified by the Wu's decision tree (Palmer and Wu, 1995). Since our aim is to see how far we can get with lexical semantics in the absence of complex representations of situational and (intersentential) discourse context, we are currently experimenting with a simple model of *default* correlations between the action expression of a sentence, and the types of the agent, instrument and action involved. We will select the action expression based on the values of these default semantic features. Similarly, we will make a default assumption about which result expression to use based on the type of the patient in the sentence. These defaults are currently implemented using the FB-LTAG feature mechanism, and hence they are not overrideable. However, we are exploring simple extensions to the FB-LTAG formalism which will allow such overrides to occur when necessary. We believe that this default lexical-semantic information can be used in many cases to determine a complete, precise compound construction without access to contextual information.

Even if no instrument is explicitly mentioned in an English source sentence, it still may be possible to make use of other lexical-semantic information to generate the correct action expression, while still avoiding the need for contextual information. We will do this by assuming that particular types of agents tend to use particular types of instruments to break things. For instance, a human being normally uses a hand, a deer uses its antlers,

INSTRUMENT	ACTION EXPRESSION
hammer, stone	*za*
axe	*kan*, *pi*
foot, hoof	*ti*
shoulder, body	*zhuang*
head, antler	*ding*
fist	*ji*, *dao*
hand	*da*

TABLE 2 Correspondences between instruments used in actions and action expressions

a horse uses its hooves.[6] This default correlation between agents and instruments can similarly be encoded in a table, so that, in the absence of an explicit instrumental adjunct, the default *break* instrument may still be chosen based on the type of the agent (which presumably will be explicitly mentioned).

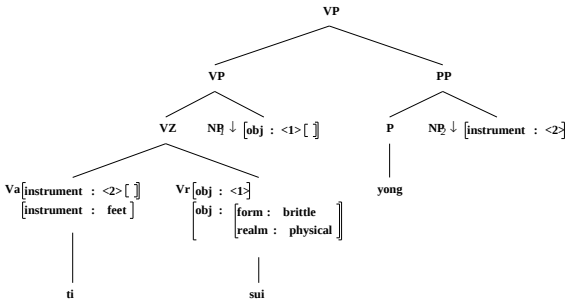
Once we have determined the selectional restrictions that define potential Chinese expressions, we can add semantic features to each lexical item, with the corresponding features on the elementary trees of the grammars. The lexical item for each Chinese verb specifies in its features what semantic restrictions it places on its object and any instrumental adjuncts which may occur. [7] The same is true on the English side where we differentiate between different senses of the lexical item *break*, which are distinguished by the object of the clause, i.e. functional break vs. physical break (Palmer and Polguère, 1994). These senses have different syntactic behaviors in English. Critically, though, the distinctions necessary for English are not forced onto the Chinese breaking verbs (or vice versa). Each noun also specifies its semantic categories, at the granularity that is necessary for this particular language. For instance, *sui* takes an object that is a physical object and is brittle, while *po* takes a solid object, as illustrated in Figure 6. The noun *chuanghu* (window) is, among other things, a physical, brittle[8] object, while the noun *men* (door) is a solid object. The corresponding noun phrase trees are shown in Figure 7. To choose the correct

[6]For the purposes of selecting an action expression in Chinese, the linguistic distinction which is sometimes made between instruments and body parts does not seem relevant.

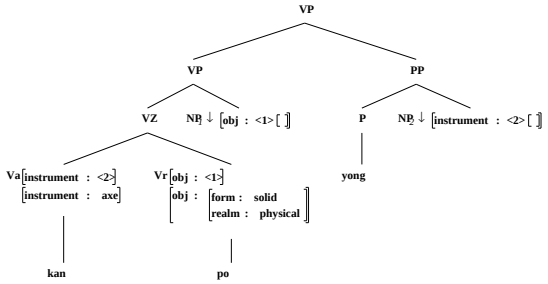
[7]Although adjuncts are not within the extended domain of locality of a verb in the FB-LTAG formalism, selectional constraints between a verb and such an adjunct may still be enforced at run-time because the features of an adjunct tree must unify with the features of the verb tree into which it is inserted.

[8]A window can be either brittle or solid, depending on various factors such as the quality of glass and the size of the window, etc. We are inclined to view such variation as contextual, and hence we exclude it from consideration for the time being. Of course, explicit adjectival modifiers such as the adjective *solid* can contribute context-independent semantic information that will override the default feature value of *brittle*.

Chinese translation for *break*, the features for the action expression and the instrument must be consistent, as shown in Figure 8.



ti'sui tree



kan'po tree

FIGURE 6 Two trees corresponding to English *break*

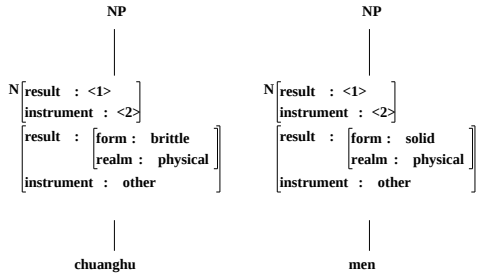
When translating the English sentence *A horse broke the window with his hooves*, the instrument *his hooves* will select for the action expression *ti*, the patient *the window* will select for the result expression *sui*, so the whole translation will be:

(7) A horse broke the window with his hooves.
yi'pi ma ti'sui na'ge chuang'hu yong ta'da ti'zi
yi'pi ma ti'sui le chuang'hu.

Similarly, the word *break* in the sentence *he broke the door with an axe* will be translated into *kan'po*, since a door is a solid object and it would select for the result expression *po*, while the instrument *the axe* would select for the action expression *kan*.

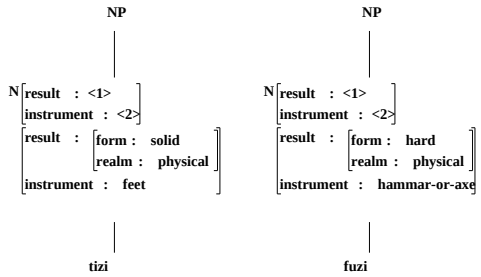
(8) He broke the door with an axe.
ta kan'po zhi'shan men yong yi'zhi fu'zi
ta yong fu'zi kan'po le men.

Sometimes, an instrument corresponds to several action expressions,



chuang'hu tree *men* tree

FIGURE 7 NP trees for Chinese *window* and *door*



ti'zi tree *fu'zi* tree

FIGURE 8 NP trees for Chinese *hoof* and *axe*

and the exact result of the action may be unclear. For example, we can use an axe to *kan* (to move horizontally) or *pi* (to move vertically). Similarly, a type of patient such as a window may be compatible with more than one result expression, since a window can be broken into small pieces, which is *sui*, or into large pieces, which is *po*. This type of ambiguity points up the limitations of our context-independent approach, and requires an interface to contextual information, a capability our system does not have.

Unlike an animate agent, a natural force doesn't use an instrument to *break* an object. Each kind of natural force has its own power and manner for exerting force on a patient. Similarly to the relationship between animate agents and default instruments, we can examine the possibility of building a mapping from natural forces to action expressions. For instance, in sentence (9), the earthquake is a series of elastic waves in the crust of the earth and it shakes the windows until the latter breaks into pieces. The Chinese verb *zhen* reflects the process of shaking, and is the correct word for an action expression. In sentence (10), it is the weight of the snow that

breaks the branch. The Chinese verb *ya* means press, and can be used to describe the snow's action.

(9) The earthquake broke every window in the house.
zhi'ci di'zhen zhen'sui mei'shan chuang'hu li zhe'ge wu'zi
zhi'ci di'zhen zhen'sui le wu'zi li de mei'shan chuang'hu

(10) The snow broke the branch.
zhi'chang xue ya'duan zhi'gen shu'zhi
xue ya'duan le shu'zhi

5.1 Comparison with a unification-based approach

The ACQUILEX system (Copestake and Sanfilippo, 1993, Briscoe, 1994) provides a typed-feature structure framework for doing MT, based on a HPSG/categorial grammar formalism for the source and target languages. Like an LTAG-based MT approach, the ACQUILEX MT framework uses a set of bi-directional transfer rules, called *tlinks*, to pair up translation equivalents in the two languages. The *tlinks* pair feature structures from the source and target languages. A translation is performed by parsing an input sentence with a unification-based source-language parser, resulting in a source-language feature structure that is mapped to a target-language feature structure, which is used to generate the output.

Since the linked feature structures may be partially parameterized, the ACQUILEX system, like the LTAG-based system, is able to make generalizations about the translations of semantically similar sentences, such as the directed-motion sentences discussed above.

While the two approaches share many of the same underlying assumptions about lexical semantic representation, the LTAG approach offers some unique advantages stemming from its use of tree composition operations. Its efficient handling of non-compositional phrases such as idioms is particularly important in the context of MT, where it is undesirable to treat such phrases too rigidly. For example, in translating the phrase *take unfair advantage of* from English, it should be possible to recognize that the target form is systematically related to the translation of the phrase *take advantage of*. An LTAG representation of these two English phrases highlights their similarity in such a way that it is easy to make reference to the common elements of both phrases when defining a transfer lexicon (Abeillé, 1990). This is not the case for the representation of these phrases in other grammars, such as the categorial grammars used by ACQUILEX.

6 Future Work and Conclusion

This work is initial work on a problem of Machine Translation that has often been ignored or relegated to 'pragmatics' or 'world knowledge'. As such, there remains much more work to be done, from extending our implementation described here to include a larger set of lexical items, to defining

ontologies for the languages that we are interested in, to questions such as how much and what kind of information is really language-specfic.

With CoGenTex, Inc, we have implemented a hybrid system with a similar transfer approach. We used the English SuperTagger (Srinivas, 1997) to do the analysis, and the CoGenTex RealPro generation system for French based on a dependency grammar to generate the French translation. The transfer lexicon was built jointly and closely resembles in capabilities the transfer lexicon presented here. We were able to semi-automatically build our transfer lexicon based on the output of Melamed's Sable system (Melamed, 1996) which automatically induces word-to-word mappings from parallel text (Nasr *et al.*, 1997). Since our domain was limited to military messages, selecting the necessary semantics features was not onerous. Building more general purpose ontologies will be much more difficult, and a suitable empirical methodology does not yet exist. Recent explorations of cross-linguistic semantic components provide promising avenues that may lend themselves to statistical methods (Dang *et al.*, 1998). Unless we are claiming that no features need to be shared between language translation pairs, which we are not, a decision must still be made about what features should be transferred between the languages - a question that must also be answered by interlingua developers.

References

Anne Abeillé, Yves Schabes, and Aravind K. Joshi. Using lexicalized tags for machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING '90)*, Helsinki, Finland, 1990.

Anne Abeillé. Lexical Constraints on Syntactics Rules in a Tree Adjoining Grammar. In 28^{th} *Meeting of the Association for Computational Linguistics*, 1990.

Ted Briscoe. Prospects for Practical Parsing of Unrestricted Text: Robust Statistical Parsing Techniques. In *Corpus-based Research into Language*. Rodopi, 1994.

J.C. Carbonell, R.E. Cullingford, and A.V. Gershman. Steps toward knowledge-based machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:376–392, 1981.

Ann Copestake and Antonio Sanfilippo. Multilingual lexical representation. In *Proceedings of the AAAI Spring Symposium: Building Lexicons for Machine Translation*, Stanford, California, 1993.

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. Investigating regular sense extensions based on intersective levin classes. In *Proceedings of ACL98*, Montreal, CA, August 1998.

B. Dorr and M.Palmer. Building an lcs-based lexicon in tags. In *Workshop on the Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity, AAAI Spring Symposium Series*, 1995.

Bonnie Jean Dorr. *Machine Translation: A View from the Lexicon*. MIT Press, Cambridge, Mass, 1993.

D. Egedi, M.Palmer, H.S. Park, and A. Joshi. Korean to english translation using synchronous tags. In *In the First Conference of the Association for Machine Translation in the Americas*, 1994.

Aravind K. Joshi, L. Levy, and M. Takahashi. Tree Adjunct Grammars. *Journal of Computer and System Sciences*, 1975.

Christian Leclerc. Organisation du lexique-grammaire des verbes francais. In *Langue Français: Dictionnaires Électroniques du Français*. Larousse, September 1990.

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

I. Melamed. Inducing bilingual lexicons. In *Proceedings of AMTA-96*, Montreal, Quebec, October 1996.

T. Mitamura. *The Hierarchical Organization for Predicate Frames for Interpretive Mapping in Natural Language Processing*. PhD thesis, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, 1989.

Alexis Nasr, Owen Rambow, Martha Palmer, and Joseph Rosenzweig. Enriching lexical transfer with cross-linguistic semantic features. In *Proceedings of the Interlingua Workshop at the MT Summit*, San Diego, California, October 1997.

S. Nirenburg, J. Carbonell, M. Tomita, and K. Goodman. *Machine translation: a knowledge-based approach*. Morgan Kaufmann, San Mateo, California, USA, 1992.

Martha Palmer and Alain Polguère. A lexical and conceptual analysis of BREAK. In *Lexical Computational Semantics*. Cambridge University Press, 1994.

Martha Palmer and Joseph Rosenzweig. Capturing motion verb generalizations with synchronous tags. In *Proceedings of AMTA-96*, Montreal, Quebec, October 1996.

Martha Palmer and Zhibiao Wu. Verb semantics for english-chinese translation. *Machine Translation*, 9:1–32, 1995.

Martha Palmer, Rebecca Passonneau, Carl Weir, and Tom Finin. The KERNEL text understanding system. *Artificial Intelligence*, 63:17–68, 1993.

Clifton Pye. Breaking concepts: Constraining predicate argument structure. Presented at the Kansas Linguistics Workshop, Lawrence, Kansas, USA, 1993.

Yves Schabes. *Mathematical and Computational Aspects of Lexicalized Grammars*. PhD thesis, Computer Science Department, University of Pennsylvania, 1990.

Stuart Shieber and Yves Schabes. Synchronous Tree Adjoining Grammars. In *Proceedings of the* 13^{th} *International Conference on Computational Linguistics (COLING'90)*, Helsinki, Finland, 1990.

B. Srinivas. Performance Evaluation of Supertagging for Partial Parsing. In *Proceedings of Fifth International Workshop on Parsing Technology*, Boston, USA, September 1997.

Jiping Sun. Interlingua-based MT through synchronous TAG. In *International Workshop on NLU and AI*, Fukuoka, Japan, 1992.

Bernard Vauquois and Christian Boitet. Automated translation at Grenoble University. *Computational Linguistics*, 11, Number 1, 1985.

K. Vijay-Shanker and Aravind K. Joshi. Unification Based Tree Adjoining Grammars. In J. Wedekind, editor, *Unification-based Grammars*. MIT Press, Cambridge, Massachusetts, 1991.

Hiroshi Yasuhara. Conceptual transfer in an interlingua method and example based MT. In *Proceedings of the Natural Language Pacific Rim Symposium (NLPRS '93)*, Fukuoka, Japan, December 1993.