Capturing Language Specific Constraints on Lexical Selection with Feature-Based Lexicalized Tree-Adjoining Grammars

Chunghye Han
Department of Linguistics
University of Pennsylvania
Philadelphia PA 19104-6228

Fei Xia, Martha Palmer, Joseph Rosenzweig
Department of Computer Science
University of Pennsylvania
Philadelphia PA 19104-6389

chunghye/fxia/mpalmer/josephr@linc.cis.upenn.edu

Abstract

The success of a Machine Translation (MT) application depends on its ability to perform *lexical selection*, that is, to choose lexical items in the target language that most closely match the lexical items in the input source. This task is particularly difficult in cases, such as those which arise in translating from English to Chinese and Korean, where the target language imposes lexical constraints which are non-existent or completely different in the source. We present an implementation of an English-Korean MT system using Feature-Based, Lexicalized Tree-Adjoining Grammar (FB-LTAG), and demonstrate its ability to handle difficulties involving lexical selection between those two languages. We also describe the applicability of this approach to similar issues which arise in English-Chinese translation. By building language-dependent FB-LTAGs for each language and then linking them via a Synchronous Tree-Adjoining Grammar (STAG), we are able to elegantly model the specific and language-dependent syntactic and semantic distinctions necessary to filter the choice of target lexical items.

1 Introduction

A Machine Translation (MT) application must be able to choose among possible translations for lexical items in its input. For example, if the system is given an English sentence with the main verb *break*, there will be many potential lexical items in Chinese which could be used to translate this verb (Palmer and Wu, , 1995). These Chinese translations for *break* will be closely related semantically and conceptually, yet distinguishing among them is critical for achieving a correct translation. Similarly, when translating the English verb *receive* into Korean, there are several alternatives to choose from, but some of these may be inappropriate depending on more global semantic factors. Determining the appropriate target lexical item to use in such cases is an important component of *lexical selection* for MT.

Previous transfer-based and interlingual approaches to MT[1] have limitations when it comes to dealing with the issues of lexical selection discussed above. For transfer-based approaches to MT such as 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. Such systems have difficulty in choosing among closely related candidate translations and cannot easily incorporate more global semantic information to make the selection. Also, these systems are not well equipped to provide a general solution for differences in syntactic realization between languages. An alternative is the interlingua approach, exemplified by Princitran (Dorr, , 1993) or Translator (Nirenburg *et al.*, , 1992), in which the source verb is mapped to a canonical semantic representation which is shared by all target languages. The elements of the semantic representation are used to select the lexical realization in each target language. While such systems are able to handle structural divergences, the construction of the large language-universal semantic ontology necessary for the interlingua is a daunting and potentially intractable task.

In this paper, we outline a proposal to capture the distinctions between the meaning of a lexical item in one language and its counterpart(s) in another language based on separate semantic-feature ontologies for each individual language. Then we show that the problem of lexical selection in machine trans-

[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.

lation can be addressed by using feature structures in the source and target grammar. The framework within which these ontologies are developed is that of Feature-Based Lexicalized Tree Adjoining Grammar (FB-LTAG) (Joshi *et al.*, , 1975; Schabes, , 1990; Vijay-Shanker and Joshi, , 1991). The grammar encodes semantic as well as syntactic features and feature constraints. We present a system for machine translation between English and Korean which is implemented in the Synchronous Tree Adjoining Grammar (STAG) formalism (Shieber and Schabes, , 1990; Abeillé *et al.*, , 1990), an extension of FB-LTAGs.

The present paper is organized as follows. In section 2, we define the problem of lexical selection in more detail. In section 3, we introduce the basic definitions of FB-LTAG and STAG and propose a model of lexical mapping between source and target languages. In section 4, we report on how the proposed model is being implemented using specific examples. Finally, section 5 discusses additional examples which pose more difficult problems.

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 *break* into Chinese.

(1) *He* *broke* *the* *door* .
ta ta'po zhi'shan men
ta da'po le men.

(2) *He* *broke* *the* *vase* .
ta da'sui zhi'ge hua'ping
ta da'sui le zhi'ge hua'ping.

As shown in Sentences (1) and (2), *break* can be translated into *da po* and *da sui* respectively, according to the physical properties of the objects that are broken. In English, *break* is a very general verb indicating an entire set of breaking events which can be distinguished by the resulting state of the object being broken. The verbs *shatter*, *snap*, *split*, etc. can all be seen as denoting 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 the selectional restrictions of *break* in English(Palmer and Wu, , 1995).

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

(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 a *request* or a *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 a *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.

3 Proposed Model

3.1 Formalism

In this paper, we present a system for machine translation between Korean and English which is implemented in the Synchronous Tree Adjoining Grammar formalism (STAG), an extension of Feature Based-Lexicalized Tree Adjoining Grammars (FB-LTAG).

FB-LTAG is based on Tree Adjoining Grammar (TAG) formalism developed by Joshi, Levy, and Takahashi (Joshi *et al.*, , 1975). An important characteristic of FB-LTAG is that it is lexicalized. That is, each lexical item is anchored to a tree structure that encodes subcategorization information. For instance, a transitive verb is anchored to a tree that

[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.

includes a subject NP and an object NP. Another characteristic of FB-LTAG is that it is feature-based. Selectional restrictions can be expressed by using the feature mechanism. The primitive elements of the standard TAG formalism are elementary trees which are of two types: initial trees and auxiliary trees. Initial trees are minimal linguistic structures that contain no recursion. They represent the phrasal structures of simple sentences. In initial trees, all internal nodes are labeled by non-terminals and all leaf nodes are labeled by terminals or non-terminal nodes marked for substitution ($\downarrow$). Auxiliary trees represent recursive structures, which are adjuncts to basic structures, such as adverbials, adjectivals and so forth. In auxiliary trees, all internal nodes are labeled by non-terminals, all leaf nodes are labeled by terminals or by non-terminal nodes marked for substitution, except for exactly one non-terminal node, called the foot node ($*$). The foot node must have the same label as the root node of the tree.

The FB-LTAG formalism defines two operations: substitution and adjunction. In the substitution operation, a node marked for substitution in an elementary tree is replaced by another elementary tree whose root label is the same as the substitution-marked node. In an adjunction operation, an auxiliary tree is inserted into an initial tree. The root and the foot nodes of the auxiliary tree must match the node label at which the auxiliary tree adjoins. Parsing of a sentence is successful only when the features in each node of the tree unify. Examples of elementary trees and the substitution operation being performed on these trees are given in Figure 1.

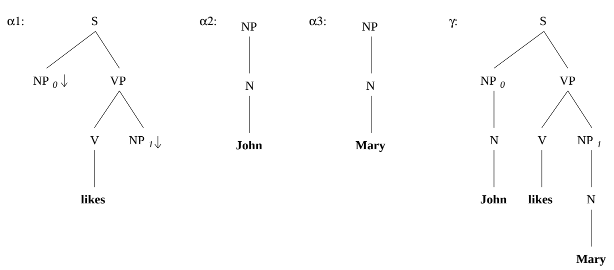


Figure 1: Elementary trees for *likes*, *John*, *Mary* and the derived tree for *John likes Mary*

In Figure 1, $\alpha 1$, $\alpha 2$ and $\alpha 3$ are elementary trees for the transitive verb *likes*, and the proper noun *John* and *Mary* respectively. After $\alpha 2$ and $\alpha 3$ substitute at NP_0 and NP_1 respectively, we get the derived tree γ for the sentence *John likes Mary*.

Introduced by (Shieber and Schabes, , 1990), STAGs are a variant of TAGs that characterize correspondences between tree adjoining languages. They can be used for relating TAGs for two different languages for the purpose of machine translation, or for relating a syntactic TAG and a semantic one for the same language for the purpose of generation. The strategy adopted for machine translation consists of matching the source FB-LTAG derivation of the source sentence to a target FB-LTAG derivation by looking at 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 tree is transferred to a target derivation by mapping each elementary tree in the source derivation tree to a tree in the target derivation tree. This is done by looking in the transfer lexicon. Finally, the target sentence is generated from the target derivation 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.

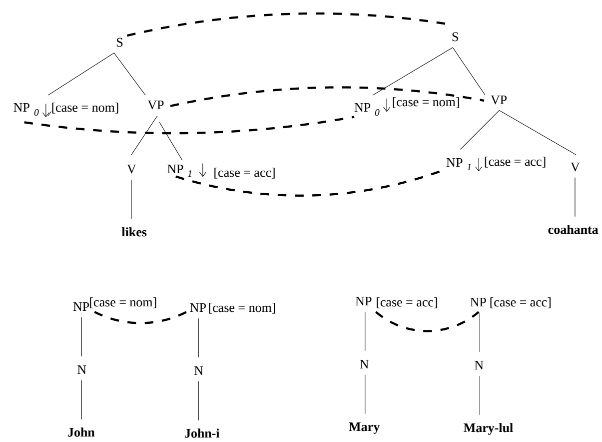


Figure 2: Lexicalized Synchronous trees for *like*, *John*, *Mary*

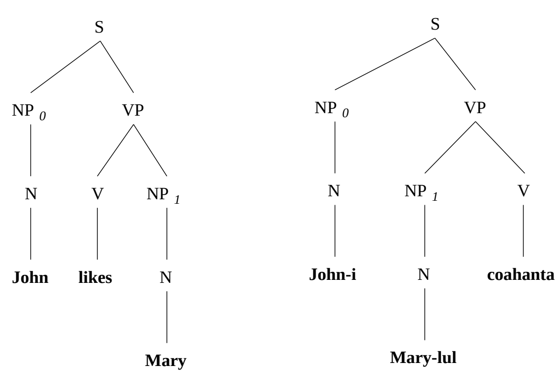


Figure 3: *John likes Mary* translated into *John-i Mary-lul coahanta*

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

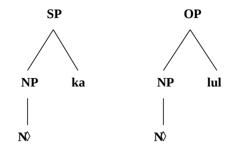
3.2 Semantic Features

We propose that semantic features and semantic constraints be specified in the usual method for each language. Since the feature unification of features and feature constraints 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.

We 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-specific 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-specific 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.

3.3 Korean TAG with Features

Egedi et. al (1994) describes the use of STAG for translating from Korean to English. They used the Korean TAG that was developed by Hyun Suk Park (Park, , 1993) which does not make use of the feature mechanism. The lack of features results in the grammar containing many unnecessary elementary trees. For instance, due to the lack of case features, a noun maps onto several different initial trees depending on which case marker it carries. That is, a noun is anchored onto an α-ka if it carries nominative case marker, onto an α-lul if it carries accusative case marker, and onto a α-eykey if it carries a dative case marker.

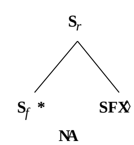


α-ka tree α-lul tree

Figure 4: Park's Noun Trees

This results in linguistically unmotivated representations such as SP (subject phrase), OP (object phrase) and DP (dative phrase) in the structural description of nouns.

Moreover, the lack of inflectional features for the encoding of sentence modes forces the grammar to contain an auxiliary tree called β-suffix that anchors mode markers on verbs.



βSUFFIX tree

Figure 5: Park's Sentential Mode Tree

The derived tree using Park's Korean TAG for the sentence in (5) is shown in figure 6:

(5) *uli-ka* *pokose-lul* *pat-ass-ta*
we-Nom report-Acc receive-Past-IND
We received the report.

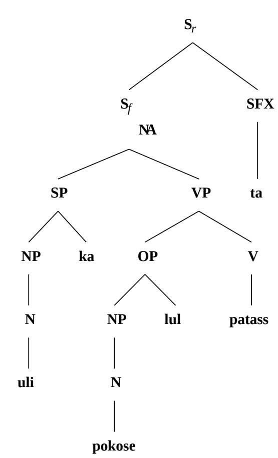


Figure 6: Park's Tree for *We received the report*

Substitution of SP *ulika* and OP *pokoselul* take place in the elementary tree that anchors the verb

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

patass. Then adjunction of βSUFFIX to the S *ulika pokoselul patass* takes place to derive *ulika pokoselul patassta*.

We have modified the Korean grammar developed by Park in such a way that it encodes semantic as well as syntactic features for each lexical item. The syntactic features include case features such as *nominative*, and *accusative*, tense features such as *past* and *present*, and mode features such as *indicative*, *imperative*, and *interrogative*. The semantic features characterize the semantic type of each lexical item. These features include such features as *human*, *process*, *information*, *object*, *location*, etc. 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, the features and feature constraints must unify. For instance, a noun is specified with case features and a semantic type feature. A verb is specified with tense and mode features and feature constraints such as case constraints and semantic type constraints on the subcategorized arguments.

The derived tree using our grammar for the sentence in (5) is shown in figure 7:

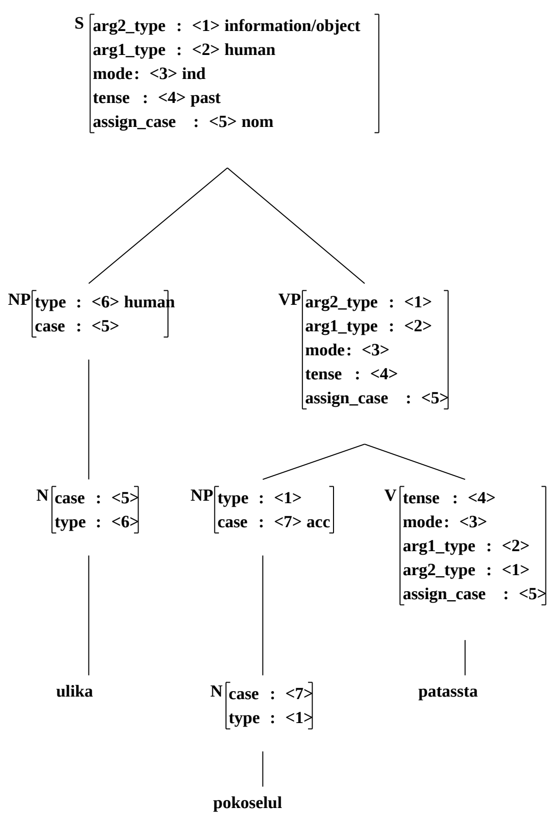


Figure 7: Our Tree for *We received the report*

The elementary tree for the verb *patassta* is encoded with features in such a way that it can only take subjects with nominative case feature and human semantic feature. Moreover, it can only take objects with accusative case feature and information or (physical) object semantic feature. The noun *ulika* is specified with nominative case feature and human semantic feature, and the noun *pokoselul* is specified with accusative case feature and (physical) object semantic feature. Since the features of the verb and the substituted nouns match, the parser accepts the sentence in (5) and the tree in figure 7 is derived.

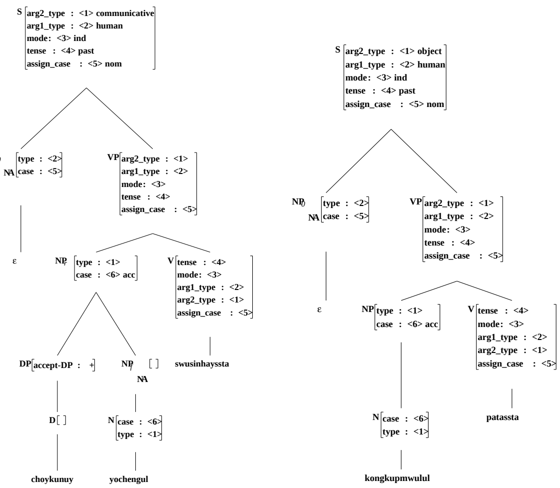
By adding features in the Korean TAG, we were able to get rid of many unnecessary and linguistically unmotivated elementary trees from the grammar. For instance, addition of case features such as nominative, accusative and dative to the grammar enabled us to make all nouns anchor onto one type of tree, αNP, resulting in one representation for nouns, NP. Moreover, by adding inflectional features such as tense and mode features, we were able to get rid of βSUFFIX tree from the grammar, resulting in a simpler structural description for sentences. And more importantly, addition of semantic features provided us with a solution to the problem of lexical selection in machine translation. We present how this is done in detail in the following sections.

4 Implementation

The proposed model 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. At this exercise, 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.

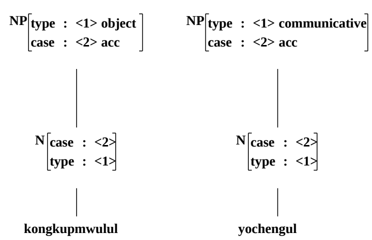
The trees in Figure 8 show the derived trees for the sentences in (3) and (4) above. The trees in Figure 9 show the NP trees for the argument NPs used in these sentences.

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



swusinhata tree *patassta* tree

Figure 8: Two trees corresponding to English sentences with *receive*



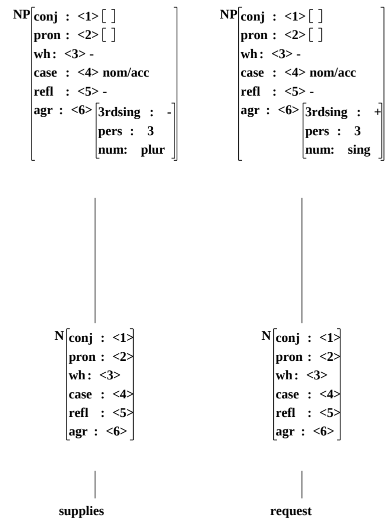
supplies tree *request* tree

Figure 9: NP trees for Korean *supplies* and *request*

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 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 10) in English therefore are not marked for their objecthood nor for their ability to be transmitted over a communicative device.



supplies tree *request* tree

Figure 10: NP trees for English *supplies* and *request*

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* (6) — 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 (7).[5]

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

[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*.

(7) *poko-lul* *swusinha-yss-ta*
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.

5 More complex examples

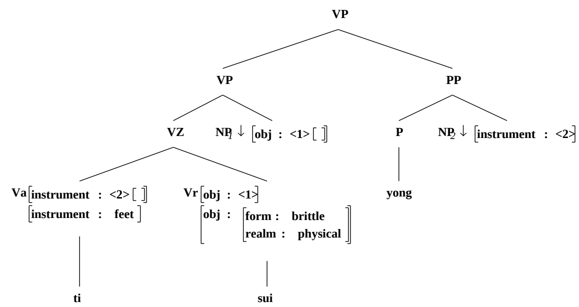
The first section briefly described the difficulties in automatically translating *break* from English to Chinese, primarily because *break* in English is a very general word, while its counterpart in Chinese needs to be more semantically precise. Not only does Chinese make more explicit the resulting state of the *broken* object, whether it is in small pieces, or pieces shaped like line segments, etc., but it also makes explicit the action that resulted in the *breaking* event, such as hitting or shouldering. The Chinese translation for *break* normally consists of two morphemes: the first morpheme (action expression) describes how the agent exerts force on the patient, the second morpheme (result expression) gives the consequence of the action.

For the result expression, there are dozens of Chinese words which describe the state *broken*. The attributes of the broken object will decide which result is most likely to occur. For example, a long, slender object such as a stick can be broken into line segments, a brittle object like a vase can be broken into small pieces, more solid objects like windows and doors can be broken into large pieces. The correct lexical choice for the result expression can often be made based on inherent characteristics of the object.

Determining the action expression is more difficult and often depends on contextual factors that may not be available to a machine translation system. 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. 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. Otherwise, we may assume 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, a horse uses its hooves.[6]

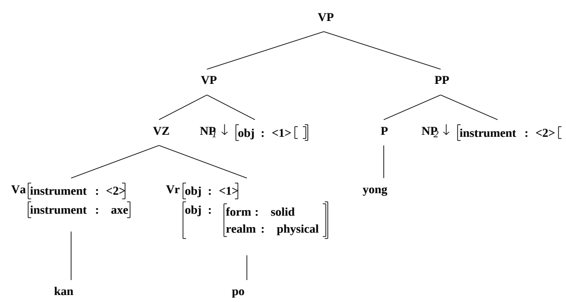
Once we have determined the selectional restrictions that define potential Chinese expressions, we

[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.



ti'sui tree

Figure 11: tisui tree corresponding to English *break*



kan'po tree

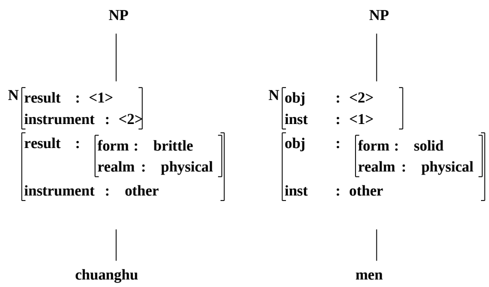
Figure 12: kanpo tree corresponding to English *break*

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] 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 and 11 and 12. 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 13.

[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.

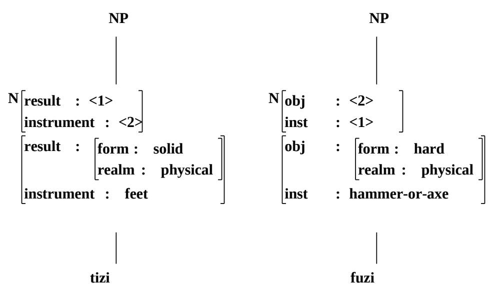
[8]A window can be either brittle or solid, depending on various factors such as the quality of glass and the size of the window, etc. We are inclined to view such variation as contextual, and hence we exclude it from consideration for the time being. Of course, explicit adjectival modifiers such as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

To choose the correct Chinese translation for *break*, the features for the action expression and the instrument must be consistent, as shown in Figure 14.



chuang'hu tree *men* tree

Figure 13: NP trees for Chinese *window* and *door*



ti'zi tree *fu'zi* tree

Figure 14: NP trees for Chinese *hoof* and *axe*

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:

(8) *A* *horse* *broke* *the* *window* *with*
yi'pi ma ti'sui na'ge chuang'hu yong
his *hooves.*
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*.

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

Sometimes, an instrument corresponds to several action expressions, 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.

The agent of the *break* action can be a natural force as well as an animate 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 a detailed discussion, please refer to (Egedi *et al.*, , 1996).

6 Future Work and Conclusion

This work is initial work on a problem of Machine Translation that has often been ignored or relegated to the domain of pragmatics or world knowledge. As such, there remains much more work to be done, from extending our implementation 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-specific. The ontologies will require access to information from many sources, not the least of which will be statistical information from aligned bilingual corpora (Gale and Church, , 1991; Dagan *et al.*, , 1993; Wu, , 1995; Wu and Xia, , 1994). 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 information should be transferred between the languages. A related question arises for interlingua approaches — what information should be included in the underlying semantic representation. It is not at all clear to us where that line should be drawn.

7 Acknowledgments

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. We are indebted to Aravind Joshi for his support of the research and the whole Xtag group at University of

Pennsylvania for their valuable input. Dania Egedi and Zhibiao Wu also provided valuable comments.

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.

I. Dagan, W. Gale, and K. W. Church. Robust word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 1993.

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.

Dania Egedi, Chunghye Han, Fei Xia, Martha Palmer, and Joseph Rosenzweig. Constraining lexical selection across languages using tree adjoining grammars. TAG+ Workshop Proceedings, editor, Owen Rambow, CSLI volume, 1996.

W. Gale and K. W. Church. Identifying word correspondences in parallel texts. In *Proceedings of the ARPA SNL Workshop*, 1991.

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

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 Zhibiao Wu. Verb semantics for english-chinese translation. *Machine Translation*, 1995.

Hyun Seuk Park. Korean grammar using tags. Technical Report MS-CIS-94-64, CIS department, University of Pennsylvania, 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.

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.

Dekai Wu and Xuanyin Xia. Learning an english-chinese lexicon from a parallel corpus. In *In the First Conference of the Association for Machine Translation in the Americas*, 1994.

D. Wu. Grammarless extraction of phrasal translation examples from parallel texts. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, 1995.