GENERATING FUNCTION WORDS IN MACHINE TRANSLATION

何 萬 順*

摘 要

本文主旨在於探討一個以詞彙功能語法爲模式的機器翻譯系統以及其中功能詞的生成方法。我們首先介紹詞彙功能語法的基本架構和一個轉換方式的機器翻譯系統，其翻譯過程有三個步驟：分析、轉換、生成。然後我們詳細討論語法功能詞的生成在這個模式下的兩個時機和方式：一是在轉換的模組，二是在生成的部份。我們以英譯中的自動翻譯爲例子：若功能詞的訊息可得自英文則於轉換部份生成，例如中文的介詞與處所詞，然而若其訊息無法得自英文則必須於最後生成階段由中文之生成語法生成，例如類詞「隻」「匹」「件」等及所有格補語詞「的」。

Abstract

This paper explores the method of generating grammatical function words within an LFG-based machine translation system. First, we give a brief exposition of Lexical Functional Grammar (LFG) and an implementation of LFG in a transfer approach with a three-stage process: analysis-transfer-generation; then, we explore some of the issues and implications involved in the generation of function words. Examples will be drawn from the English-to-Chinese module. Prepositions and locative formatives are examples of function words generated in transfer, while classifiers and the possessive complementizer, information of which is not available in the source language, are generated in the generation component after transfer.

0. INTRODUCTION

In any given natural language, two primary, distinctive classes of words, or morphemes to be more specific, can be distinguished: lexical versus grammatical words. While lexical words, also known as content words, have relatively easily identifiable semantic content, i.e., they denote entities, events, states, properties, and the like, grammatical words, also known as function words, may instead simply express grammatical relations or functions within a sentence (e.g., Kaplan 1989: 74-75). What are often referred to as major or open classes, such as verb, noun, adjective, and adverb, usually contain lexical words, and the so-called minor or closed categories, for example presposition, classifier, particle, and conjunction etc., are largely composed of function words. Since the function categories usually came from grammaticalized lexical

*作者為本校語言學研究所副教授

categories (e.g., Vincent 1993, Hopper and Traugott 1993), most languages share the major categories, but the grammatical function categories tend to be less stable and are a great source of variation across languages.

The problem of translation between any two languages can be generalized as one of "mismatching" – all language systems are differently structured, semantically as well as grammatically. Yet, in this respect, the degree of variation in the overt expression of various grammatical relations is certainly far greater than that of semantic variation. Frequently, the source language contains grammatical categories not found in the target languages, or vice versa (e.g., Thunes 1994); Chinese noun classifiers, unknown in the English grammar, are a good example; furthermore, the same grammatical relations may be expressed via completely different devices. For example, English subject and object are configurationally encoded, as [NP, VP]s and [V,NP]vp respectively, most often without any overt case marking, while in Japanese the same grammatical relations are non-configurationally encoded by overt case makers *ga* and *o* respectively. Thus, in a machine translation system, while it is often possible to have a more robust procedure to translate lexical items and defaulting may produce approximate translations, to match the grammatical relations requires more delicate maneuvering and defaulting may produce no result at all.

In the first generation systems before the 1970's, machine translation, or MT in short, was viewed largely as an engineering task in its nature rather than a linguistic one; current MT researchers and system developers, however, having recognized the severe limitations of previous methodology, have come to place linguistic analysis as the primary task and therefore in general base the overall design of the translation system on a certain form of contemporary linguistic theory (e.g., Slocum 1985, Hutchins and Somers 1992, Nirenburg 1987). Among all the contemporary grammatical theories, the constraint-based unification grammars have been popular for the computational processing of natural languages (e.g., Shieber 1986, Levin 1990). Lexical Functional Grammar (e.g., Kaplan and Bresnan 1982, Kiparsky 1985, Sells 1985, Wescoat 1987, Kaplan 1989), or LFG, in particular, has received much attention in its suitability as the linguistic model for either transfer-based or interlingua-based machine translation systems (e.g., Arnold et al 1990, Thunes 1994, Her et at 1991a, 1994, Kudo and Nomura 1986, Kaplan et al 1989, Dyvik 1992). Her (1994), in particular, argues for the suitability of LFG theory and the modified formalism employed in the system described here for natural language processing applications.

The system depicted here has been developed by Dan Higinbotham, Joseph Pentheroudakis, and One-Soon Her; the overall design philosophy can be found in Higinbotham (1990b), Pentheroudakis (1990a), and Her et al (1991a). Several language-

paris have been developed. The English-Chinese system is, to some extent, described in Her (1990) and Her et al (1994). The English-Korean bi-directional system, developed together with Jay Kim (Kim 1988, 1991, Kim and Pentheroudakis 1991, Pentheroudakis 1990b), is now implemented in the TACCIMS system employed by the Joint Forces in South Korea. The MANTRA Project in University of Bergen has been developing English-Norwegian bi-directional system (Thunes 1994). The English-Japanese system has been delivered to an international translation organization in Japan and there is also an English-Arabic bi-directional system under development in Jordan.

In this paper we first describe a transfer-based MT system modeled after the current grammatical theory of Lexical Functional Grammar and illustrate its three stages of automatic translation process: analysis, transfer, and generation. Then, we discuss some of the issues and implications involved in the generation of function words within such a system. Examples will be drawn from the English-to-Chinese module, although the translation engine itself may be adopted to other languages, as it has already been done with Japanese, Korean, Arabic, and Norwegian. Chinese prepositions and locative formatives are used as examples of function words generated in the transfer stage, and classifiers, information of which is available only in the target language, and the possessive complementizer *de* are used to demonstrate the generation of function words in the generation component by the generation grammar.

1. LFG AND THE TRANSFER APPROACH TO MT

Current implementation of the 'indirect' translation approach in MT consists of three steps: analysis, transfer or interlingua, and generation, as shown in Fig. 1 below. The translation process is 'indirect' in that the analysis of the source sentence is not directly affected by the consideration of the target language. In fact, ideally the analysis of the source language and the generation of the target language are motivated completely independent of each other, and the interface between analysis and generation is either a transfer method or interlingua method (e.g., Slocum 1985, Her et al 1991).

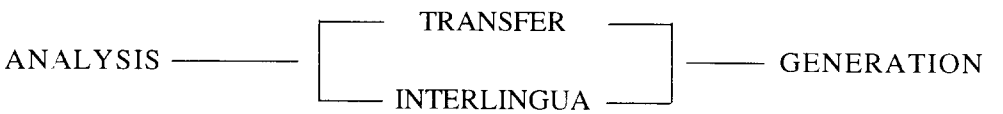


Fig. 1. Indirect Approach of MT

The system to be depicated here opts for the transfer method[1]. Within this methodology, the transfer module takes the representation of the correct analysis of the source sentence as its input and transforms it into an approximation of that of the desired target sentence (e.g., Tucker 1987). The generation component then provides the transformed representation with the rest of the necessary information and linearizes all the relevant lexical items into the target sentence. More specifically, since the system employs LFG as its linguistic framework, the result of the independent analysis of a source sentence is a tree representation of its constituent structure, or c-structure, as well as a corresponding bracket graph representation of functional structure, or f-structure. An example is given below in Fig. 2.

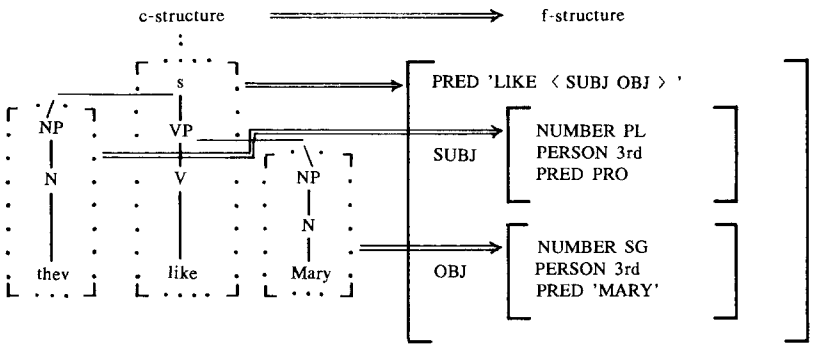


Fig. 2. Co-description of c- and f-structure

The most important design feature of LFG is the distinction of the external structure and the internal structure. The external structure, which varies greatly across languages, expresses precedence, dominance, and categories; the formal model of this is the c-structure, represented as a phrase structure tree. The internal structure, the formal model of which is f-structure, represents the way meaningful grammatical functions are grouped into semantic predicate argument relations, and this structure remains largely invariant across languages. These two structures, like the lyrics and the melody of a song, mutually describe or constrain each other and are related not by isomorphy but by local co-description (Bresnan 1993). Within this framework, the proper analysis of a phrase or a sentence is thus a c-structure together with a corresponding f-structure. And the f-structure, given its relative cross-linguistic invariance serves ideally as the basis for transfer.

Within the LFG-based machine translation system, the tree representation of the c-structure is discarded once analysis is complete, because the highly language-dependent c-structure information is inconsequential to carry over to the target sentence. Furthermore, all the necessary information on the meaning and grammatical relations

are preserved within the f-structure, which is a non-linear hierarchical attribute-value matrix representing the relatively language-universal underlying grammatical relations within a sentence and therefore serves ideally as the basis for transfer among different languages. The ordering of attributes on the same level of f-structure is random. Therefore, it is easily demonstrateable that the transfer of f-structure is much more straightforward than that of the constituent (tree) structure. That is one of the significant characteristics which sets this current system apart with other English-Chinese MT systems such as the BehaviorTran system developed by Tsing Hua University and BTC Center (Su et al 1992), the NTUMT system at the National Taiwan University (Lin 1986, Chui et al 1989), and the MITTRAN system of Matsushita Electric Institute of Technology, Taipei, all of which are tree-transfer-based; ERSO's TransMaster system, based on the case grammar, however, is closer to our system here in that the transfer component is based on an abstract representation of the internal structure, rather than the external constituent structure. Incidentally, none of the systems above, except the relative success of BehaviorTran, has reached the mature stage of actual practical production of usable translation.

Since the system depicted here is based on LFG, during the transfer and generation stages, what is being transformed is the graph representation of the f-structure. Fig. 3 illustrates the flow of operations during the execution of the translation engine. More detailed operations during transfer and generation are given.

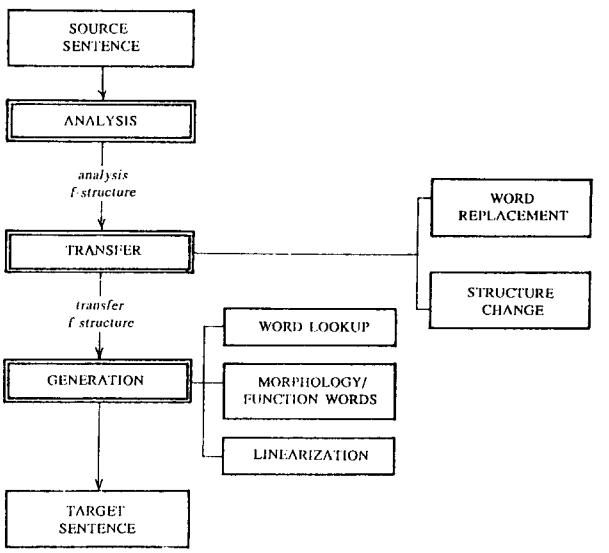


Fig. 3. Translation procedure and modules

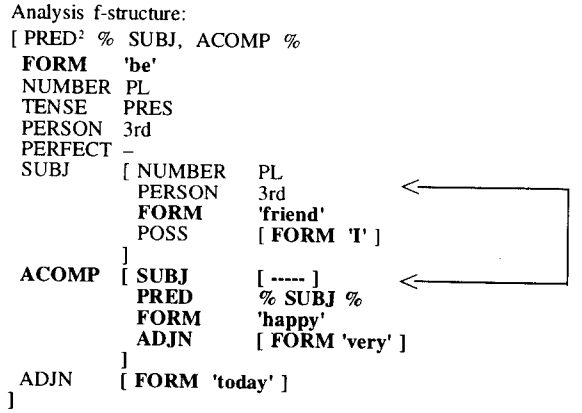
Lexical replacement (i.e., translation of lexical words) as well as necessary structural changes on the analysis f-structure are performed according to transfer rules evoked by lexical transfer entries. The output of the transfer operations is the transfer f-structure, which in turn serves as the basis for generation.

Note that, since the system is transfer-based, we do not treat f-structure as the so-called interlingua, which is claimed to be a language-independent universal representation of meaning (see note 1); rather, the f-structure of the source sentence is still language-dependent and serves only as the basis for transfer to become the f-structure of the target language. (What we do claim, however, is the transfer of a tree structure is much more cumbersome than that of the functional structure.) Thus, all linguistic entries and rules must be language-dependent (with language names specifically indicated, for example in 2 and 3 below "eR' denotes English rule, "ec' specifies an English-Chinese transfer entry, "cW' is a Chinese word entry, etc.), while the translation engine itself is language independent. (As mentioned earlier, more than eight language pairs have been developed under the same translation engine.) Within the transfer method, it is thus necessary to have a source-target dictionary and a specific set of transfer rules for any specific pair of languages with a fixed direction of translation. The generation module proceeds in three stages: first, target words in the transfer f-structure are looked up in the target lexicon and information unified with the transfer f-structure; second, necessary inflection of morphological elements and target function words are added; and finally, the linearization rules traverse the generation f-structure and arrange the specified lexical items into the target sentence string. In 1a-e, we show a simple example of the five stages of this translation process of the sentence *my friends are very happy today.*

1.a. Source sentence:

My friends are very happy today.

1.b. Analysis f-structure:



1.c. Transfer f-structure:

```
[ PRED  %  SUBJ  %
  FORM     'gaolxing4'              'lexical replacement
  NUMBER   PL                       'structural change
  TENSE    PRES                     'ACOMP is absorbed
  PERSON   3rd                      'into the next upper level
  PERFECT  -
  SUBJ     [ NUMBER    PL
             PERSON    3rd
             FORM      'peng2you3'
             POSS      [ FORM  'wo3' ]
           ]
  ADJN     { [ FORM  'jin1tian1' ]
             [ FORM  'hen3' ]
           }
]
```

Notice that in the transfer f-structure all English lexical words are replaced with Chinese translations and furthermore the entire ACOMP in the analysis f-transfer is absorbed into the immediate higher level after the transfer operation, because in Chinese the predicative adjective, or stative verbs to be more accurate, need not be introduced by linking verbs such as *be*. The transfer f-structure then enters the generation module, which produces generation f-structure 1d and finally the output target sentence 1e.

1.d. Generation f-structure:

```
[ PRED  %  SUBJ  %
  FORM     'gaolxing 4'
  INFLFORM[3]  'gao1xing4'                  'inflected form
  NUMBER   PL
  TENSE    PRES
  PERSON   3rd
  PERFECT  -
  SUBJ     [ NUMBER      PL
             PERSON      3rd
             FORM        'peng2you3'
             INFLFORM    'peng2you3men'     'inflected form
             POSS        [ FORM  'wo3'
                           INFLFORM 'wo3'
                           POSSFORM 'de'    'function word
           ]               ]
  ADJN     { [ FORM  'jinitian1'
               INFLFORM  'jin1tian1'
               TYPE  TIME ]
             [ FORM  'hen3'
               INFLFORM  'hen3'
               TYPE  DEGREE ]
           }
]                                           'linearization  ↓
```

1. e. Target sentence:

Wo3 de peng1you3men jin1tian1 hen3 gao1xing4.

During the generation process, *peng2you3* is inflected with *men* as INFLFORM *peng2you3men*, while other forms remain unchanged; meanwhile, the function word *de* is added (as POSSFORM) to the POSS subsidiary f-structure. All information thus far still cannot produce the desired target sentence. Linearization rules for Chinese must correctly arrange the various subsidiary f-structures and forms contained within the generation f-structure. This illustrates the modular transfer approach to machine translation.

Note that the transfer stage can in fact be viewed as the first step towards generation, in the sense that all transfer operations, like those in the generation stage, are for the specific purpose to create a correct f-structure of the intended target sentence. Therefore, as we will illustrate with the generation of Chinese function words in the context of English-to-Chinese, the linguist sometimes needs to make a conscious decision as to when to perform a certain operation, i.e., during transfer or generation, although often the linguistics of the language pair does dictate one way or the other.

2. LINGUISTIC ORGANIZATION

There are basically two types of linguistic records in the system: lexical entries and rules. For any given language pair in translation, all lexical entries are organized into three sublexicons: source, transfer, and target; take the English-to-Chinese system as an example, there are English word entries, English-Chinese transfer entries, and Chinese word entries; examples are given in 2a-c respectively.

```
2.a. eW_friend ::                    'English word
     [ CAT  N                        'source lexical entry
       FS   [ FORM   'friend'
              HUMAN+
            ]
     ]

2.b. ec_friend ::                    'English to Chinese
     [ FS   [ FORM   'peng2you3' ]   'transfer lexical entry
     ]

2.c. cW_peng2you3 ::                 'Chinese word
     [ CAT  N                        'target lexical entry
       FS   [ FORM 'peng2you3'
              HUMAN  +
              CLASS  'wei4'
            ]
     ]
```

Similarly, rules must belong to a specific module: analysis, transfer, or generation. Analysis rules, such as 3a, are responsible for assigning the correct c- and f-structure to the source sentence, transfer rules, e.g., 3b, transform the analysis f-structure to approximate the desired f-structure of the target sentence, and finally generation rules generate the correct translation. Example 3a builds category S and its corresponding f-structure; note also that it puts the functional information of NP into the SUBJ subsidiary f-structure (see Fig. 2). 3b, on the other hand, is responsible for the structure change in transfer, as shown in 1c earlier.

```
3.a. eR_S ::                          'English rule; rule name
     [ NP : SUBJ                      'info in NP mapped to SUBJ
       VP
     ]
     →
     BUILD (S)                        '= S → NP  VP

3.b. eR_STD-TRAN-PRED-ADJ ::          'Eng-Chi transfer; rule name
     [ IF < ↑ FORM > = c 'be'         'If  FORM is 'be'
          &&                          'and
          < ↑ ACOMP >                 'ACOMP is found
       THEN  ABSORB ( ↑ , < ↑ ACOMP > 'see Note 4
     ]
```

Generation rules in turn are composed of 1) inflection (or function word) rules, e.g., 3c, which give the target lexical items appropriate inflections and provide the necessary target function words, and 2) linearization rules, e.g., 3d, which traverse the final generation f-structure, pick out all the existing specified lexical forms, and linearize them in specified word order in the target language.

```
3.c. eR_FW-NOUN ::                    'Chinese inflection and
     [ NOUN                           'function word rule for N
     ]
     →
     (IF  < ↑ HUMAN > = c +           'must be human
          &&
          < ↑ NUMBER > = c PL         'must be plural
          &&
          < ↑ QUANTIFIERS > =NONE 'must not find QUANTIFIERS
          THEN < ↑ INFLFORM > = '< ↑ FORM > men'
     )
     &&
     •
     •
     &&
     < ↑ GENERATION-RULE > = cR_LINEAR-NP  'apply cR_LINEAR-NP
```

```
3.d. eR_LINEAR-NP ::          'linearization rule for NPs
     [ •
       •
       SPFORM                 'the determiner
       QUANTIFIERS            'quantifiers
       CLASSFORM              'the classifier
       INFLFORM               'the head noun
       •
       •
     ]
     →
     SUCCEED ( )    'output string in the order: SPFORM > (string formed
                    'within) QUANTIFIERS > CLASSFORM > INFLFORM
```

Each source entry has a corresponding transfer entry, which, along with the transfer rule(s) it specifies, governs the lexical selection and the manner of f-structure transformation. Each lexical selection in transfer should constitute a target entry with information necessary for the generation of inflection, function words, and linearization. Notice that the function word rule 3c, where plural human nouns are inflected with the morpheme *men*, and it can also be expanded to insert function words such as the ordinal marker *di4* for numerals, possessive complementizer *de*, and aspect makers *le, zhe, guo4*, etc. Note that this rule also specifies, at the end, that the linearization rule cR_LINEAR-NP (3d) should be evoked to arrange the lexical items and grammatical functions on its level. 3d can, of course, be further expanded to include more relevant lexical items contained within the Chinese NP.

3. GENERATION OF CHINESE FUNCTION WORDS

Chinese grammarians have made the distinction between 'full' and 'empty' words since the tenth century. This distinction is equivalent to the more contemporary distinction between 'content' and 'function' words. A lexical unit which is empty of lexical meaning and assumes a grammatical function is known as a 'function word'; prepositions are typical examples. Chinese, as a textbook case of isolating languages, has little inflectional morphology; therefore, besides word order, a rich set of function words are employed to encode grammatical functions. Hence, the correct translation or generation of function words plays an important role in the English-to-Chinese module. Certain English function words do find counterparts in the target language; in cases like this, the translation process is straightforward as that of the lexical replacement of content words. Nonetheless, in most cases they have to be added according to the grammatical functions they play. If such information is available in the source analysis

f-structure, then the corresponding target function words may be added either during transfer or until the generation stage; however, if the grammatical information is target-specific and thus not available until after the target word lookup, then obviously it would be impractical to contemplate on the correct function words during transfer.

3.1 Function Words in Transfer

Certain grammatical functions in English are expressed through function words and have a straightforward correspondence in Chinese. Like content words, such function words are translated in the transfer component according to their transfer entries. Most of prepositions in English correspond with Chinese prepositions quite straightforwardly. However, while location in English is expressed through various prepositions such as *in, on, above*, etc, in Chinese it is encoded with the preposition *zai4* together with a postpositional locative formative (see 4-7 below)[5].

4. a. *on* the car
   b. zai4 che1 shang4
5. a. *in* the car
   b. *zai4* che1 *li*
6. a. under the car
   b. *zai4* che1 *xia4*
7. a. *by* the car
   b. *zai4 che1 pang2*

Therefore, what needs to be accomplished during transfer is to map the various English locational prepositions to the Chinese preposition *zai4* while at the same time insert the correct corresponding postpositional locative formative. We will use 5a as an example. The f-structure of 5a is shown in 8a. 8b shows the English-Chinese transfer entry of *in*, in which it is specified that the Chinese form to replace *in* is *li* and also that the transfer rule named STD-TRAN-LOC-PP should be evoked.

8.a. Analysis f-structure of 'in the car'

```
[ PFORM  'in'
  FORM  'car'
  DEFINITE +
  NUMBER  SG
  USE  # VEHICLE #
]
```

8.b. ec_in : :

```
[                                   'transfer lexical entry of 'in'
  FS [ FORM 'li' ]                  'in is to be replaced with li
  \ STD-TRAN-LOC-PP                 'evoke this transfer rule
]
```

```
8.c. ec _STD-TRAN-LOC-PP : :        'transfer rule
   [ IF ~ < ↑ IDIOMATIC > =c +   'must not be idiomatic
     THEN    TRAN ( ↑ , PFORM)                    '1
             &&
             REPLABEL ( ↑ , PFORM, LOCFORM)       '2
             &&
             < ↑ PFORM > = 'zai4'                 '3
   ]
```

The transfer rule STD-TRAN-LOC-PP specifies that if the string involved is not marked as idiomatic and thus does not require an idiomatic translation, then 1) TRAN translates the English prepositional form PFORM (i.e., *in*) to the form specified in the transfer entry (*i.e., li*), 2) the operation REPLABEL replaces the label PFORM with locative form LOCFORM, and 3) the label-value pair [PFORM 'zai4'] be assigned to the current f-structure[6]. The result of the transfer rule is shown in 8d, which also shows that *car* has been replaced with *che1*.

```
8.d. Transfer f-structure of 'in the car'
   [ PFORM  'zai4'
     LOCFORM  'li'
     FORM   'che1'
     DEFINITE  +
     NUMBER  SG
     USE  # VEHICLE #
   ]
```

Then, this transfer f-structure enters the generation phase where *che1* as a target word is looked up in the target lexicon and information in its lexical entry, 8e, is unified with the transfer f-structure. After inflection or function word insertion, the system inserts an INFLFORM, which has as its value the inflected FORM. In this case, FORM and INFLFORM are identical, since no inflection applies. The result is the generation f-structure shown in 8f.

```
8.e. eW_che1  : :                    'Chinese word entry of che1
   [ CAT  N
     FS [ CLASS  'liang4' ]          'liang4'
          FORM 'che1'                'che1'
        ]
   [

8.f. [ PFORM 'zai4'                  'generation f-structure of 'in the car'
     LOCFORM 'li'
     FORM 'che1'
     CLASS  'liang4'
     INFLFORM  'che1'
     DEFINITE  +
     NUMBER  SG
     USE  # VEHICLE #
     GENERATION-RULE cR_LINEAR-NP
   ]
```

Notice that in the generation f-structure there is a GENERATION-RULE label whose value indicates which linearization rule governs the ordering of lexical elements and grammatical functions on this level of the f-structure. In this case the linearization rule specified is cR_LINEAR-NP, shown below in 8g. The transfer operations together with generation operations thus correctly generate the Chinese string *zai4 che1 li* from *in the car.*

```
8.g. eR_LINEAR-NP  ::          'linearization rule for NPs
     [ •
       •
       PFORM                   'the preposition, zai4
       SPFORM                  'the determiner
       POSS                    'possessive NP
       QUANTIFIERS             'quantifiers
       CLASSFORM               'the classifier
       INFLFORM                'the noun, che1
       LOCFORM                 'the locative, li
       •
       •
     ]
     →
     SUCCEED ( )
```

3.2 Function Words in Generation

Although in natural languages grammatical functions constitute a closed and limited inventory, rarely two languages share the exact same set of grammatical functions, unless of course they are intimately related. Noun classes encoded by classifiers are such a case. The classification of nouns by classifiers is common among Oriental languages but only scarcely found in European languages. Furthermore, even between two languages that do employ classifiers for noun class distinction, for example Chinese and Japanese, the classifications encoded by classifiers do not correspond neatly. In an English-to-Chinese MT system, information on such target-language-specific grammatical encoding must be specified in the individual target lexical entries; see 9a-b below. The examples of 10 and 11 illustrate the lack of consistency and regularity of noun classes by classifiers in Chinese and thus the fact that such information cannot be predicted through the semantic features of the nouns, contrary to claims made by some cognitive grammarians, such as Lakoff (1987, 1991) and Lee (1988).

```
9.a.. cW_niu2  : :              'target lexical entry
     [ CAT  N                   'of niu2 'cow'
       FS  [ FORM 'niu2'
             ANIMATE +
             CLASS  'tiao2'     'noun class
           ]
     ]
```

```
9.b. cW_ma3 ::                    'target lexical entry
     [CAT N                       'of ma3 'horse'
      FS    [ FORM 'ma3'
              ANIMATE +
              CLASS 'pil'         'noun class
            ]
     ]

10.a. sanl tiao2 niu2             'three cows'
   b. sanl tiao2 xian4            'three lines'
   c. sanl pil ma3                'three horses'
   d.*sanl tiao2 ma3              'three horses'

11.a. sanl zhangl zhi3            'three pieces of paper'
   b. sanl zhangl zuolzi          'three tables'
   c. sanl ba3 yi3zi              'three chairs'
   d.*sanl zhangl yi3zi           'three chairs'
```

Consequently, the correct generation of classifiers can only be accomplished during the generation phase after lexical look-up in the target lexicon, for only then is the noun class information unified into the transfer f-structure. 12a below is the analysis f-structure of the phrase *three horses*, 12b the transfer f-structure, and 12c the f-structure in generation after unification with the lexical entry of *ma3*, i.e., 9b above.

```
12.a. Analysis f-structure of 'three horses'
      [ DEFINITE    -
        FORM 'horse'
        ANIMATE     +
        NUMBER PL
        QUANTIFIERS {[ FORM 'three' ]}
      ]

12.b. Transfer f-structure of 'three horses'
      [ DEFINITE    -
        FORM 'ma3'
        ANIMATE     +
        NUMBER PL
        QUANTIFIERS {[ FORM 'sanl' ]}

12.c. In generation, after unification with entry of ma3 (9b)
      [ DEFINITE    -
        FORM 'ma3'
        CLASS 'pil'
        ANIMATE +
        NUMBER PL
        QUANTIFIERS {[ FORM 'sanl' ]}
      ]
```

However, although the information on noun class is specified in the noun entries, the classifier is not allowed unless there is a modifying quantifier or determiner. See the contrast between 13a and 13b below. The correct classifier is required in 13a, while barred in 13b. 13c indicates 13b's f-structure halfway through, that is, it is after unification with entry of *ma3* (9b) but before function word insertion.

13.a. sanl pil ma3 'three horses'
b. wo3 de (*pil) ma3 'my horse'

13.c. *Wo3 de ma3* 'my horse'; in generation, after unification with entry of *ma3* (9b) but before function word insertion

```
[ POSS    [ FORM 'wo3'          'wo3 de ma3 'my horse'
          ]
| DEFINITE     -
  FORM 'ma3'
  CLASS 'pi1'  -
  AMIMATE      +
  DEFINITE     -
  NUMBER PL
  QUANTIFIERS {[ FORM 'sanl' ]}
]
```

We thus have to make sure that the classifier only gets inserted when there is a quantifier present in the f-structure. This is accomplished in the function word insertion rule, cR_FW-NOUN, shown in 14a below, which is specified in the noun entry *ma3* (9b) and evoked at lexical lookup. This rule copies the value of CLASS to the label CLASSFORM only when the label QUANTIFIERS is found. In addition, it also inserts the complementizer *de* when the noun is possessive and thus is contained in POSS. Finally the rule specifies that the linearlization rule cR_LINEAR-NP is to govern the ordering of lexical elements on this level of f-structure.

```
14.a. eR_FW-NOUN ::                'inflection/
   [ NOUN                          'function word rule
   ]
   —
   ( IF ( ↑ , QUANTIFIERS )
     THEN < ↑ INFLFORM > = '< ↑ FORM > men'
   )
   &&
   ( IF < ↑ HUMAN > =c +           'must be human
        &&
        < ↑ NUMBER > =c PL         'must be plural
        &&
        < ↑ QUANTIFIERS > = NONE   'must not find QUANTIFIERS
     THEN < ↑ INFLFORM > = '< ↑ FORM> men' 'inflected with men
   )
   &&
   ( IF ( ↑ ↑ , POSS)   'if contained with POSS
     THEN < ↑ POSSFORM > = 'de'
   )
   •
   •
   &&
   < ↑ GENERATION-RULE > = cR_LINEAR-NP
```

After the operations in this function word rule, the f-structure of 12c will now have a correct classifier (CLASSFORM) inserted. After all inflection (function word) rules have applied, we have the generation f-structure of *three horses* (14b). As for 13c, *my horse* will not get a classifier due to the absence of QUANTIFIERS, see 14c, which also observes the account where *de* as a complementizer marks the following nominal element as the head of the construction (Ross1981).

14.b. In generation, after function word insertion

```
[ DEFINITE    -
  FORM 'ma3'
  INFLFORM    'ma3'
  CLASSFORM   'pi1'
  ANIMATE +
  DEFINITE -
  NUMBER PL
  QUANTIFIERS {[ INFLFORM 'sanl' ]
                 FORM 'sanl'
               ]}
  GENERATION-RULE cR_LINEAR-NP
]
```

14.c. In generation, after function word insertion

```
[ POSS [ FORM 'wo3'
         INFLFORM     'ma3'
         POSSFORM     'de'
         ]
  FORM 'ma3'
  INFLORM 'ma3'
  CLASS 'pil'
  DEFINITE -
  GENERATION-RULES cR_LINEAR-NP
```

The specified linearization rule then will traverse the f-structure and arrange lexical items and grammatical functions as specified in the rules. We will repeat 8g, the partial linearization rule for noun phrases, as 15a below.

15.a. cR_LINEAR-NP :: 'linearization rule for NPs

```
[ •
  •
  PFORM          'the preposition
  SPFORM         'the determiner
  POSS           'possessive NP, e.g., wo3 de
  QUANTIFIERS    'quantifiers, e.g., snal
  CLASSFORM      'the classifier, e.g., pil
  INFLFORM       'the noun, e.g., ma3
  LOCFORM        'the locative
  POSSFORM
  •
  •
]
→
SUCCEED ( )
```

With the information of noun class specified in the noun target entries and through the operations of lexical lookup, function word insertion, and finally the linearization rule, we are able to generate the correct strings *sanl pil ma3* (QUANTIFIERS) > CLASSFORM > INFLFORM), *wo3 de ma3* ([INFLFORM > POSSFORM]$_{poss}$ > INFLFORM), and also *wo3 de sanl pil ma3* ([INFLFORM > POSSFORM]$_{poss}$ > QUANTIFIERS > CLASSFORM > INFLFORM). The same scheme works for more complex constituents.

4. CONCLUDING REMARKS

In this paper, we have shown that the system based on the theory of Lexical Functional Grammar is expressive and flexible in generating Chinese function words. When the lexical selection can be based on information of the source language, function words can be inserted during the transfer stage, such is the case with Chinese locational preposition *zai4* and postpositional locative formatives. However, if such selection is dependent upon information in the source lexicon, then the correct function word must be inserted via inflection rules during generation; noun classifiers in Chinese are such a case. Furthermore, since the transfer component is intimately tied to a particular language pair and is thus the most ad hoc module, it is better to specify function word insertion in the generation grammar whenever possible to avoid repeating similar transfer operations in each language pair. In so doing the generation grammar can be made more portable to other language pairs; the generation of Chinese possessive complementizer *de* is provided as an example.

NOTES

*Two anonymous reviewers have made valuable suggestions to improve the paper; I thank them for their keen observations.

1. As stated in Her et al (1991:286-7), 'given the current state of art of AI research an adequate, suitable interlingua is not yet possible and also that a totally language/culture-independent interlingua may not exist...in order to build a practical translation system, it is best to start off with a shallower analysis that may provide a relatively language-independent representation and deepen the analysis as experience and theory advance.' For the description of interlingua-based MT systems, see Carbonell and Tomita (1987), Nirengurg (1989), or Dyvik (1992), and for a comparison of an LFG transfer implementation and an interlingua system, see Thunes (1994).
2. The PRED feature in conventional LFG formalism consists of the semantic form and the subcategorization frame of a predicate, for example 'devour <SUBJ OBJ>'; we have separated the PRED feature into two features in our formalism: PRED, that specifies the subcategorization of grammatical functions such as % SUBJ OBJ%, and FORM, which specifies its semantic form such as 'devour'. Therefore, non-predicative elements, for example 'floor', will only have FORM 'floor', but not the PRED feature. This modification is for the convenience of generating morphologically-inflected forms; see note 3 below.

3. The value of FORM will be transformed into that of INFLFORM by the morphological operations specified in the generation grammar. For example, FORM 'book'will create INFLFORM 'books' if on the same level there is NUMBER PL. It is the value 'books' of INFLFORM that will be picked out as part of the translation output. Also, see rule 14 for an example in Chinese.
4. From the operations indicated in 1a-c, it can be seen that the ABSORB function must take two arguments, source followed by target, for example ABSORB (↑ , < ↑ ACOMP>), and the system also defines the function as 'unify the source level with the target level and let all information in the source level be the default'; in other words, ABSORB does not fail in cases of value conflicts, instead the feature-value pair in the source level will be preserved while the conflicting value of the target level will be overridden.
5. While zai4's prepositional status is generally fairly non-controversial (e.g., Chao 1965, Lu 1984, Her 1990), the analysis of the locative formatives such as shang4, li, and wai4 is of some uncertainty. These locative formatives are known as 'relator nouns' in the analysis where they are considered heads of the NP construction (e.g., Starosta 1985, 1988). However, in our treatment here, following Her (1990), we deem the noun the head, while the locative formative contributes the feature [PLACE +] to the NP construction and thus satisfies zai4's subcategorization requirement of a place noun.
6. As one anonymous review pointed out, it is of course favorable to use monotonic lexical replacement whenever possible, as it is declarative and thus more efficient. However, procedural operations like the current one may be necessary as lexical replacement is not always straightforward, for example, the mapping between English 'on' and Chinese 'zai4...shang4'. In addition, structural changes may also demand procedural operations.

REFERENCES

Arnold, D., I. Crookston, L. Sadler, and A. Way, 1990. LFG and Translation, in *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation*, Linguistic Research Center, Austin, Texas.

Bresnan, J., (Ed.), 1982. *The Mental Representation of Grammatical Relations,* MIT Press, Cambridge, Mass.

Bresnan, J. 1993. Lexical-Functional Grammar: Lecture Notes and Problem Sets, Linguistic Department, Stanford University.

Carbonell J. and M. Tomita. 1987. Knowledge-based Machine Translation, the CMU Approach. In R. Nirenburg, (Ed.). 1987. 68-89.

Chui, K. et al., 1989. NTUMT strategy for prepositional-phrase attachment, Proceedings of ROCLING II (1989).

Dyvik, H., 1992. Semantics-based Machine Translation and the Concept of "Translation Equivalence", paper presented at the 12th Scandinavian Conference of Linguistics. August 17-21 1993, Gothenburg.

Her, O. 1990. Grammatical Functions and Verb Subcategorization in Mandarin Chinese. Crane Publishing, Co., Taipei.

Her, O. 1994. Application of grammatical theories and formalisms in natural language processing, *Research Newsletter of the National Chengchi University*, January 1994, Vol. 2, 109-27.

Her, O., D. Higinbotham, and J. Pentheroudakis, 1989. An LFG-based English-Chinese Machine Translation System. *Proceedings of 1989 International Symposium on Chinese Text Processing* 8.3-7. Boca Raton: Florida Atlantic University.

Her, O., D. Higinbotham, and J. Pentheroudakis, 1991a. An LFG-based Machine Translation System, *Computer Processing of Chinese and Oriental Languages*, Vol. 5, Numbers 3 & 4, 285-297.

Her, O., D. Higinbotham, and J. Pentheroudakis, 1991b. The Treatment of Idioms in the LFG-based ECS Machine Translation System, *ACH/ALLC 1991 Conference Handbook*, Dan Ross and David Brink (Eds.) 191-196. Tempe: Arizona State University.

Her, O., D. Higinbotham, and J. Pentheroudakis., 1994. Lexical and Idiomatic Transfer in Machine Translation: An LFG approach, in *Research in Humanities Computing 3,* Susan Hocky and Nancy Ide (Eds.), Oxford: Oxford University Press, 200-16.

Higinbotham, D., 1987. Morpho-LFG. *Procedings of the Eleventh Annual Symposium of the Deseret Language and Linguistics Society*. Provo: Brigham Young University.

Higinbotham, D., 1990a. Semantic Co-occurrence Networks and Automatie Resolution of Lexical Ambiguity in Machine Translation. Ph.D. Dissertation. University of Texas at Austin.

Higinbotham, D., 1990b. Digital Thought Mapping: Automatic Translation of Natural Language. *Proceedings of the Topical Meeting on Advances in Human Factors Research on Man/Computer Interactions: Nuclear and Beyand* 365-369. American Nuclear Society.

Higinbotham, D., 1990c. Semantic Co-occurrence Networks. *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.* Austin: Universityof Texas.

Higinbotham, D., 1991. The Resolution of Lexical Ambiguity in Machine Translation. *Proceedings of the Eleventh Annual Symposium of the Deseret Language and Linguistics Society.* Provo: Brigham Young University.

Hopper, P. and E. Traugott, 1993. *Grammaticalization,* Cambridge University Press.

Hutchins, W. and H. Somers, 1992. An Introduction to Machine Translation, London: Academic Press.

Kim, J., 1988. Parsing the Korean Verb. *Proceedings of the Sixth International Conference on Korean Linguistics*. Seoul: International Circle of Korean Linguistics.

Kim, J., 1991. *A Lexical-Functional Grammar Account of Light Verbs.* Ph.D. Dissertation. University of Hawaii.

Kim, J. and J. Pentheroudakis, 1991. English-Korean Bi-directional MT System. *The 32nd Annual Meeting of American Translators Association,* Salt Lake City.

Kaplan, J., 1989. English Grammar: Principles and Facts, Prentice Hall International Ltd.

Kaplan, R., 1989. The Formal Architecture of Lexical-Functional Grammar, in *Proceedings of ROCLING 11*. 1-18.

Kaplan, R. and J. Bresnan., 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation, in J. Bresnan (Ed.), 173-281.

Kaplan, R., K. Netter, J. Wedekind, and A. Zaenan., 1989. Translation by Structural Correspondence, in *Proceedings of the Fourth Conference of the European*

Chapter of the Association for Computational Linguistics, University of Manchester, 272-281.

Kiparsky, C., 1985. LFG Manual, Manuscript, Xerox Palo Alto Research Center, Palo Alto, California.

Kudo, I. and H. Nomura. 1986. Lexical-Functional Transfer: A Transfer Framework in a Machine Translation System based on LFG. *Proceedings of Coling 1986*, Bonn. 112-114.

Lakoff, G., *1987. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind,* University of Chicago Press, Chicago.

Lakoff, G., 1991. Cognitive versus Generative Linguistics: How Commitments Influence Results. *Language and Communication*, 11.1/2, 53-62.

Lee, M., 1988. Language, Perception, and the World. In J. Hawkins (Ed.), *Explaining Language Universals*, Basil Blackwell, Oxford.

Levin, L. 1990. Syntactic Theory and Grammar Design for Machine Translation, in *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation*, Linguistic Research Center, Austin, Texas.

Lin, L., J. Huang, K. Chen and L. Lee. 1986. A Chinese Natural Language Processing System Base upon the Theory of Empty Categories. *Proceedings AAAI-86 Fifth Conference on Artificial Intelligence*,

Volume II. 1059-1062.

Lu, S. 1984. *Xiandai Hanyu Ba Bai Ci* (Eight Hundred Words in Modern Chinese). Hong Kong: Shangwu Publishing Co.

Nagao, M., J. Tsujii, and J. Nakamura., 1985. The Japanese Government Project for Machine Translation, *Computational Linguistics* 11.2-3: 91-100.

Nirenburg, S. (Ed.). 1987. *Machine Translation: Theoretical and Methodological Issues.* Cambridge University Press.

Nirenburg, S., 1989, Knowledge-Based Machine Translation, *Machine Translation* 4.1: 5-24.

Pentheroudakis, J., 1990a. You Can Get There from Here: Design and Implementation Issues in Machine Translation Systems. *Proceedings of the Topical Meeting on Advances in Human Factors Research on Man/Computer Interactions: Nuclear and Beyond* 370-375. American Nuclear Society.

Pentheroudakis, J., 1990a. Complex Bi-diurectional Transfer in the ECS English-Korean System. *Proceedings of SICONLP '90* 108-119. Seoul: Seoul National University.

Pentheroudakis, J. and D. Higinbotham, 1991. Morfogen: A Morphology Grammar Builder and Dictionary Interface Tool. *Proceedings of Deseret Linguistic and Language Society Seventeenth Annual Symposium.* Provo: Brigham Young University.

Ross, C. 1981. On the Functions of Mandarin *de*. *Journal of Chinese Linguistics* 11 (2).

Sadler, L. and H. Thompson, 1991. Structural Non-Correspondence in Translation, in *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, 293-298.

Sells, P. 1985. *Lectures on Contemporary Syntactic Theories*. Stanford, CA: CSLI, Stanford University.

Shieber, S. 1986. *Introduction to Unifrication-based Approaches to Grammar*. Standford, CA: CSLI, Standford University.

Slocum, J., 1985. A Survey of Machine Translation: Its History, Current Status, and Future Prospects, *Computational Linguistics* 11.1: 1-17.

Starosta, S. 1985. Mandarin Case Marking. *Journal of Chinese Linguistics* 13.2:215-266.

Starosta, S. 1988. *The Case for Lexicase*. London: Pinter Publishers.

Su, K. and J. Chang. 1992. Why corpus-based statistics-oriented machine translation, Proceedings of TMI-92, pp. 249-62.

Thunes, M., 1994. *Transfer and Interlingua in Machine Translation: A Comparison of Two Implementations,* Department of Linguistics and Phonetics, University of Bergen, Bergen, Norway.

Tucker, A., 1987. Current Strategies in Machine Translation Research and Development, in r. Nirenburg, (Ed.), *Machine Translation: Theoretical and Methodological Issues*, Cambridge University Press, Cambridge, 22-41.

Vincent, N., 1993, Exaptation and Grammaticalization, to appear in *Proceedings of the International Conference of Historical Linguistics 12*, H. Anderson (Ed.), Amsterdam: John Benjamins.

Wada, H., 1990. Discourse Processing in MT: Problems in Pronominal Translation. *Proceedings of COLING-90*, 73-75.

Wescoat, M., 1987. Practical Instructions for Working with the Formalism of Lexical Functional Grammar. In J. Bresnan (Ed.). Lexical-Functional Grammar. Course Material for L1229, 1987 Linguistic Institute, Stanford University.