A Generation System of KMITT's MT Project

Ms. Kanlaya Narue-domkul
Mr. Booncharoen Sirinaovakul
Asst.Prof. Nuantip Tantisawetrat
Machine Translation Laboratory
King Mongkut's Institute of Techonology Thonburi

Abstract This paper provides an overview of a part of the machine translation project which has been studying in Machine Translation Laboratory at King Mongkut's Institute of Technology Thonburi. The project aims at developing a generation system which is to generate the Thai language as a target language in the machine translation process. In this study, the generation system produces Thai sentence from the interlingua which represents the meaning of the source language. This interlingua expresses conceptual relations by using semantic cases resulted from the analysis system. There are three main steps involves in generation system, first, syntactic generation which creates the syntactic structure of the target language by using the generation grammars; second, words selection which selects the most appropriate Thai word for each concept in the interlingua basing on word category, subcategory and other related issues; third, words ordering which orders the words according to the patterns of the Thai sentences.

The most important aspect for the generation process is the dictionary. The generation dictionary contains syntactic information, verb patterns and mapping patterns required for semantic-syntax mapping.

This system is developed and tested with 50 simple sentences. There are 189 words in the dictionary. The study has indicated that the target language (Thai) developed is accurate and reliable in representing the source language in the translation process.

1. **Introduction.** One of the famous strategies in Machine Translation(MT) system is the Interlingua MT strategy. The main idea of this strategy is that the source language and the target language never contact directly. The meaning of the source language sentence is represented in an artificial language, called 'INTERLINGUA'. The process of this translation begins with analyzing the source language sentence. The output of the analyzed sentence is represented by the interlingua which is dependent from any form of a particular lan-

guage. This interlingua represents the semantic structure of the input sentence. The target language, then, is generated directly from this interlingua. One of the advantages of using this process is that it reduces a lot of redundant information. For instance, n different languages, n concept dictionaries and n sets of *grammar rules are needed for n(n-1) translations.*

In this paper, the proposed method uses the interlingua as a strategy of generation and artificial intelligence as a technique for problem solving. The reasons are described as follows:

First, the development of the dictionary and the grammar rules of any language are related with the interlingua, not only for analyzer but also for generator. The analysis module and the generation module are connected by interlingua. Therefore, the dictionary and the grammar rules for each module can be developed separately.

Second, by using artificial intelligence technique, the grammar rules in the knowledge base can be developed independently from computer programming. Therefore, the created rules are maintainable. The knowledge developer can develop and edit the grammar rules without touching the computer program. It is convenient for the developer who is scare of computer language to develop grammar rules for the system.

This paper provides an overview of a generation system which has been studied in Machine Translation Laboratory at King Mongkut's Institute of Technology Thonburi. The generation process, the designed system and the results and conclusion showing examples of generation process are discussed in details.

2. **Generation Process.** The project aims at developing a generation system which is to generate the Thai language as a target language of the machine translation. The interlingua resulted from the analysis system is used as an input. The developed prototype is limited to a simple sentence, and each sentence is considered independently. The databases used in the process are a generation dictionary (IL-TL dictionary) and other tables which are related to the process. In this study, the generation system produces Thai sentence from the interlingua which represents the meaning of the source language. This interlingua, resulted from the analysis system, expresses the conceptual relations by using semantic cases.

The three main steps in generation process are Syntactic Generation, Words selection and Words ordering. Syntactic Generation creates the syntactic structure of the target language by using the generation grammar. Words Selection selects the most appropriate Thai word for each concept in the interlingua. Words Ordering orders the generated words.

2.1. Syntactic Generation. The interlingua, represented in a semantic tree structure, is the input of the generation system. The syntactic generation procedure is the fist procedure to process the interlingua. It creates the syntactic structure for the Thai sentence by mapping the interlingua's semantic relation with the syntactic relation. The procedure consists of dictionary loading, syntactic mapping and subject selection.

Dictionary Loading process searches the information for all conceptual primitive (CP) from generation dictionary. The process is done by #LDICT command in the knowledge base and each CP-name is used as a keyword. In searching for CP's information, if there are some CPs that have more than one set of information, the appropriate information will be selected by Syntactic Mapping procedure.

Fig.1 is the example of dictionary information for CP-name "TALK" that has two sets of information.

CP-name : TALK

CONCEPTUAL	*ENTRY*	*TCAT*	*TSUBCAT*	*TMAPS*	*TVP*	*AKO*
TALK	คุย	*V*	*V*	*SUB=AGT,COMP=OBJ*	*3*	*2111*
TALK	คุย	*V*	*V*	*SUB=AGT*	*1*	*2111*

Fig.1 Dictionary information

Syntactic Mapping procedure maps the semantic relations (only the relations between root node and its daughters) with TMAPS of the root node. At this state, the most appropriate information of the root node is selected and the syntactic cases are also mapped. For the interlingua which its relations are both obligatory and free cases, the procedure has to cut free cases by comparing them with the free-case table before selecting the syntactic cases.

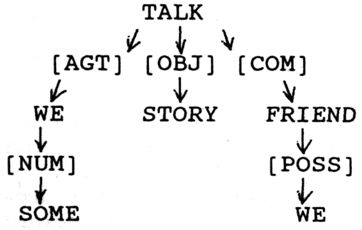


Fig.2a

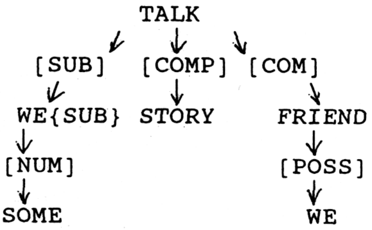


Fig.2b

Subject Selection process searches for the branch whose syntactic•case is 'SUB' and moves it to the left most branch. Then, the procedure assigns value 'SUB' to the feature of its daughter.

Fig.2a is the interlingua. It composes of two obligatory cases, AGT and OBJ, and one free case, COM. Fig.2b shows the output of the syntactic generation procedure. The value 'SUB' is assigned to Node WE.

2.2. Words Selection. Words selection is the main task of the generation system. It selects the most appropriate Thai words and maps them onto the CPs of interlingua. In this project, the process of words selection are as follows.

2.2.1. Thai word generation. The process maps Thai words onto the root node. This is done by retrieving Thai words from the "ENTRY" field of the information, selected by syntactic mapping procedure. For daughter node, The system uses the syntactic case of the syntactic structure as an information for selecting the Thai word for the CP. This syntactic case indicates category or subcategory of the daughter node. The system compares this syntactic case with the category or subcategory of the information loaded from dictionary. The Thai word that has the same category or subcategory as indicated by syntactic case will be selected.

For other nodes, or leaf nodes which their parent node is not root node, their TSUBCATs are selected by using NMAPS table. NMAPS is defined by considering the relation of meanings between noun and its modifiers. The example of NMAPS is shown in fig.3.

Semantic Case	TSUBCAT
NUM	DDBQ, DIAN, JNRN, JNRP, NCNM
CAP	NCMN
POSS	PPRS, NCMN
...	....

Fig.3. NMAPS table

From fig.3., the process compares the leaf node's case with the semantic case of NMAPS table. Then, the set of TSUBCATs of the matched semantic case is loaded. These TSUBCATs are compared with the TSUBCATs loaded from dictionary and the intersected TSUBCATs are selected.

2.2.2. Classifier Generation. In Thai, a noun has a classifier when it is numbered. The classifier node of the noun node is created when the words selection procedure is done. The procedure processes by looking at the field 'num' of the information loaded from the dictionary. The noun node that has the value in the field 'num' generates the classifier node by using its value as CP-name. Feature, named 'clas', and the case, named 'CLAS', are assigned to that classifier node. This classifier node will be revised after the next procedure, feature generation, is done and the unnecessary classifier node will be cut.

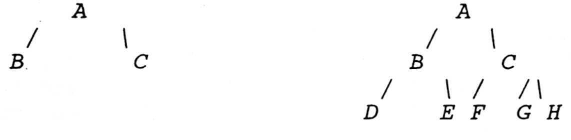
2.2.3. Feature Generation. There are some additional information of a word such as modality, speech act, which expresses the feature of CP and/or case. The representation of feature for each particular language is not the same. The feature occurred in one language may not occur in other languages. The system, therefore, has to test for the occurrence of the feature in Thai language by using Thai grammar rules. If the occurrence exists, the procedure will assign the appropriate Thai word for that feature. In assigning the Thai word, Rules, Category and Subcategory are used for consideration. After the Thai word is assigned, the new node is generated by using this Thai word as its CP-name. In this project, there are four types of case indicating the relation of new generated node, namely PREP, PLUR, NEG and TENS. The assignment of case name for the new generated node depends on the grammar rules.

The output of the syntactic generation procedure is in tree structure. Thai words are given to all CP-nodes. The information which is needed for generating Thai sentence are also attached.

2.3. Words Ordering. There are two steps in this module, linear structure creation and words ordering.

2.3.1. Linear Structure Creation changes the tree structure, from syntactic generation process, to be the linear structure. The process begins by moving the left most branch of the tree to the left side of its parent and moving the rest branches of the tree to the right side.

Example 1



[B] [A] [C] *[D] [B] [E] [A] [F] [C] [G] [H]*

(a) *(b)*

From Example 1.a, the daughter node B is moved to left side and the daughter node C is moved to right side of their parent A. Example 1.b shows linear creation of the three level tree. Considering from the lowest level, node D is moved to the left side of B and E is moved to the right side of B. Node D, B and E, then, are moved to the left side of A. The algorithm repeats the same process to the right side of A. The transformed linear structure of the example 1.b is shown under the tree structure.

2.3.2. Words Ordering is the last step in the generation process. The steps of words ordering are: Grouping, NP ordering and AUX ordering.

- Grouping divides the nodes into three groups by using verb node as a separator. These groups are NP, V and NP. The auxiliary verb nodes, that precede the verb node, are moved to the right-hand side (before verb). Such move is done by comparing two nodes at a time. After all the auxiliary verb nodes are placed before the verb. The system marks the first auxiliary verb node with 'F'. Then, it groups the auxiliary verb nodes after the verb node and marks the last auxiliary verb node with an 'L'.

- NP ordering deals with the word order in a noun phrase. The system uses the pattern of Thai noun phrases as a look-up table. The pattern is represented by transition network as shown in Fig.4 . The steps of NP ordering starts with the NP-group before F-node and, then, the NP-group after L-node.

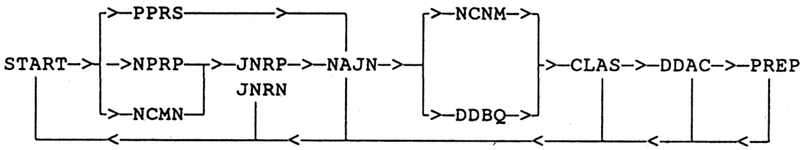


Fig.4. NP Ordering

- AUX Ordering starts from the F-node to the L-node. These nodes are ordered by using auxiliary ordering table as shown in fig.5. The Output from this process is the Thai sentence which is the target of this project.

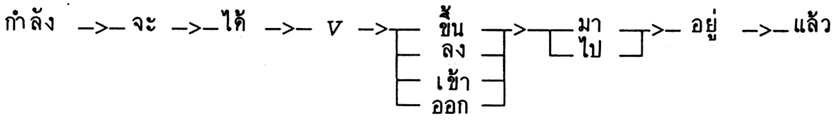


Fig.5. AUX Ordering

3. <u>Designed</u> <u>System.</u> This is based on the idea of efficiency and maintainability of generation system, a particular system has been designed by using the artificial intelligence technique. The designed system is shown in fig.6.

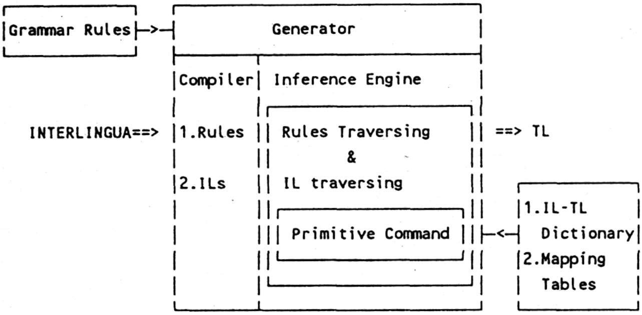


Fig.6. The designed generation system

The system is classified into four parts.

Interlingua (IL): Interlingua which is produced by the analysis module is an input of the system.

The generator: The generator controls generation processes.

Grammar Rule: Grammar Rules are the Thai generation grammar rules.

Generation Dictionary: Generation Dictionary is the interlingua to target language dictionary.

3.1. Interlingua. So far, there has been a great deal of effort in creating an effective interlingua. A great effort has been put into the research (on how to represent the meaning of the context correctly and completely) since the accuracy of the representation of a source language by an interlingua will result in the correctness of generation of the target language.

The interlingua is a representation of the deep meaning of a source language text by a symbol. It should

- express the meaning of a context directly.
- represent the same characteristics of all languages with the same symbol.
- be independent from the particular features of a language.

In this project, the interlingua structure which is used in the process is semantic tree structure and it has its own word called CP-name and syntax called semantic case. The interlingua composes of three parts: conceptual primitive (CP), case relation (CASE) and feature(FEATURE).

CP expresses the concept of the content word of the sentence. To define a CP-name for a word, the meaning of the word must be considered. The different words which have the same meaning when occur in the same context have the same CP-name: such as

CP-names	Thai
STORY	นิทาน
STORY	เรื่อง
THAT	ให้
THAT	ว่า
THAT	นั้น

and in the same way, the word which has several meaning must have different CP-names such as

CP-name	Thai
DESCEND	ลง
DOWN	ลง
HUMAN	คน
STIR	คน

CASE represents the relation between two concepts. Thai semantic cases are designed for this relationship representation. The definition of Thai case is considered from the relation between noun-noun, noun-verb, noun-modifier, verb-modifier, modifier-modifier, modifier-sentence and clause-clause. There are 37 Thai cases which have been studied in Machine Translation Laboratory at King Mongkut's Institute of Technology Thonburi.

FEATURE represents the information of word meaning in the sentence such as modality, speech act etc. in the interlingua.

An Interlingua structure is represented by a dependency tree structure which is composed of CP nodes, CASEs and FEATUREs.

In the interlingua, there are three types of CP nodes, namely 'root node', 'daughter node' and 'leaf node'. A root node is the top node of the tree structure. This node represents the concept of the main verb of the sentence. A daughter node is the second level of the tree structure. A leaf node is the node in the lower level than the daughter nodes of the tree structure. Fig.7 shows the three types of nodes in a tree structure.

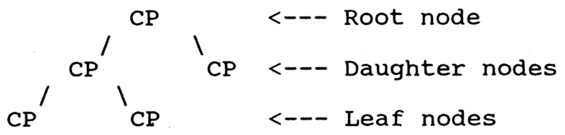


Fig 7. The tree structure.

CASE of the interlingua structure is placed in square brackets on the arc of relation between two conceptual nodes. The arrow on the arc is used for distinguishing the head and dependent node. The arrow of the arc is pointed from head to depender. The FEATURE of both CP and CASE are represented in the parenthesis. From fig 8., CP1 is the head node and CP2 is the dependent node. The relation between CP1 and CP2 are represented by CASE and the arrow is pointed from CP1 to CP2. The FEATURE is expressed in the parenthesis beside its node.

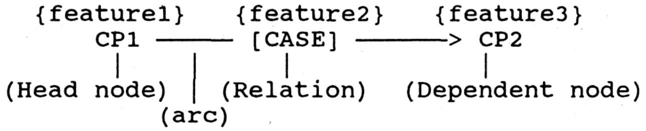


Fig 8. The relation.

Fig.9 shows the interlingua representation of an example sentence "Most Americans remembered that story since childhood.".

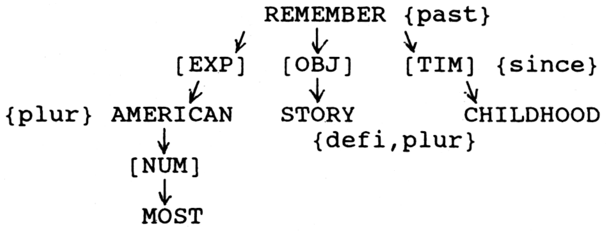


Fig.9 The interlingua.

Besides the graphical representation, the interlingua is also represented in LISP language structure. The syntax and its explanation is shown in Fig.10.. Fig.11. shows the example of interlingua generated from "Some of us are talking about the story."

```
* CP1{fea1} ( [CASE{feac}] CP2 {fea2}).
```

*   *Root node indicator.*
CP1 *Head node.*
CP2 *Dependent node.*
fea1 *Feature of CP1 node. It is represented in {} with small letter. If there is more than one feature, it will be separated by ','.*
() *Dependent node of the head node CP1. If there is more than one dependent node in the same level, it will be separated by ','.*
[] *Case between conceptual nodes. It is expressed in [ ] with capital letter.*
feac *Case feature.*
.   *The terminator.*

Fig.10. The interlingua syntax.

```
* REMEMBER{past} ( [EXP] AMERICAN{plur} ([NUM] most),
                   [OBJ] STORY {defi,plur},
                   [TIM{since}] CHILDHOOD
                 ).
```

Fig.11. The example of interlingua.

3.2. Generation Dictionary. The most important aspect of the generation process is the dictionary. The generation dictionary contains information about the categories and subcategories of the word, the syntactic information(VP) for mapping pattern and their concepts. The pattern is required for semantic-syntax mapping. The details of dictionary fields are as follows.

CONCEPTUAL	ENTRY	TCAT	TSUBCAT	TMAP	TVP	AKO	SYNONYM	NUM
AGREE	เห็นด้วย	V	V	SUB=EXP.COMP=COM	2	2112	-	-
AGREE	เห็นด้วย	V	V	SUB=EXP.COMP=OBJ	12	2111	-	-
BREAKFAST	อาหารเช้า	N	NCMN	-	-	1321	มื้อเช้า	มื้อ
FATHER	พ่อ	N	NCMN	-	-	111	บิดา	คน

CONCEPTUAL is the concept of the word. It is used as a keyword in searching for the Thai word, which has the same concept. This field is the connector between the source language and the target language. The definition of the concept has been studied in Machine Translation Laboratory at KMITT.

ENTRY is a Thai word. The following is the example of TVP table.

TCAT is Thai Category. It is a part of speech which is classified by the function of the word such as n-noun, v-verb, and etc.. This field is used for subject selection and word selection.

TSUBCAT is Thai Subcategory. It is a subcategory of part of speech such as NPRP-proper noun, NCMN-common noun, PPRS-personal pronoun,etc. This field is used for word selection and word ordering.

TMAPS is Thai mapping. It is a mapping table between syntactic case and semantic case. Only the obligatory cases of verb are mapped in the table. This field is used for mapping the semantic case onto syntactic case.

TVP is the Thai verb pattern. It indicates the kind of structures in which each verb is used. It is for selecting the Thai word for a daughter node. The value of TVP is a number which is used as an index for mapping the Thai verb pattern. There are 19 Thai verb patterns for simple sentences which have been established in this project. The examples of TVP table are shown bellows.

AKO is " A Kind Of ". It is a word meaning hierarchy which is used for Thai words selection for the concepts, the prepositions and classifiers.

SYNONYM is field for Thai words which have the same meaning.

NUM is Number. It is a classifier which occurs only for Thai nouns.

Example 2

TVP TABLE

1:	SUB,V
2:	SUB,V,PP
3:	SUB,V,COMP
4:	SUB,V,DOB
5:	SUB,V,DOB,PP
....	

3.3. Grammar Rules. Thai Grammar which is used for generating Thai sentence is represented in the format shown in fig. 12a. The '# Header' represents the label of the rule. It may or may not exist.

A rule consists of two parts : condition and action. They are represented in the parenthesis ( ). The separator between condition and action is the marker ':'. A rule may have more than one condition and more than one action.

The knowledge base structure can be divided into phases. A phase is composed of many rules. There are 35 phases in this knowledge base. The details are shown in fig.12b.

```
# Header 1
  ( Condition 1   :   Action 1 )
# Header 2
  ( Condition 2   :   Action 2
                  :   Action 2.1
                  :   Action 2.3   )
# Header 3
  ( Condition 3   :
  ( Condition 3.1 :
  ( Condition 3.2 :   Action 3.2
                  :   Action 3.3 )))
```

Fig. 12a The rules format.

Phase 1	:	Load Dictionary
Phase 2-3	:	Syntactic Generation
Phase 4	:	Subject Selection
Phase 5-6	:	NMAPS Mapping
Phase 7-12	:	Feature Creation
Phase 13-20	:	Thai words Selection
Phase 21-27	:	Feature Creation
Phase 28	:	Linear Structure Creation
Phase 29-35	:	Words Ordering

Fig. 12b The detail of knowledge base.

3.4. Generator. The generator of the system consists of two parts, Compiler and Inference Engine. Compiler is the functions for checking the knowledge base and interlingua syntax. Inference Engine is used for interlingua and rule traversing. In traversing along the interlingua, the grammar rules are also inferred. The interlingua traversing is depth first traversing. Fig.13b shows the traversing steps of the inference engine along the interlingua in Fig.13a.

```
                [1]
              /     \
        [2]             [5]
      /     \         /     \
   [3]       [4]   [6]       [7]
```

Fig.13a The interlingua tree.

[1]-->[2]-->[3]-->[2]-->[4]-->[2]-->[1]-->[5]-->[6]-->[5]-->[7]-->[1]

Fig.13b The traversing steps.

When the inference engine traverses to the last node, it will automatically goes back to the root node. However, if the moving command is used, the traversing will be controlled by that command.

While traversing, the current node can refer t other nodes by using the window. Fig.14 shows th window system markers.

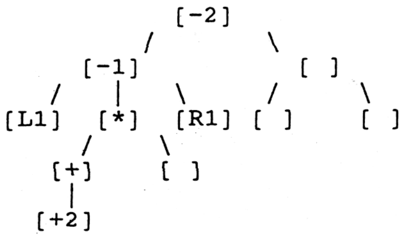


Fig.14 The window marker.

* : Current node.
-n : The nth node upper the current node.
+n : The next nth node below the current node.
Ln : The nth node to the left of the current node.
Rn : The nth node to the right of the current node

It is noted that the Ln and Rn node must be th daughters of the same head node.

The direction of rules traversing is the 'Forwar Chaining'. Fig.15 shows the rule traversing steps o the rules shown in fig.12.

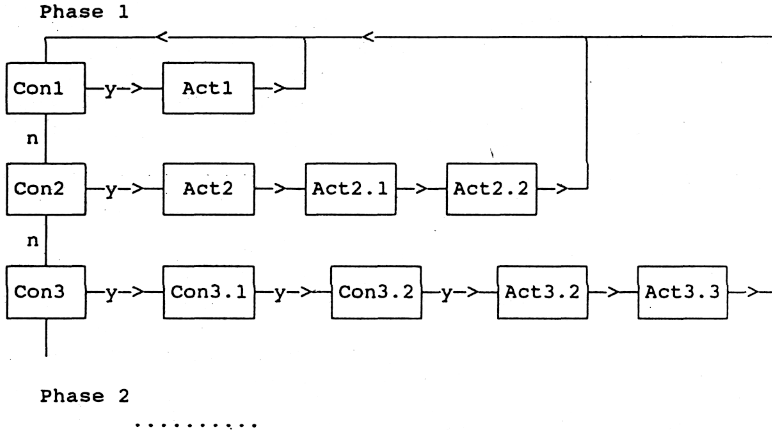


Fig.15 The algorithm of rule traversing

From Fig.15, the inference engine starts by chec ing the first condition "Con1". If the condition i satisfied then the inference engine takes action "Act1" and goes back to the first condition or else the inference engine checks the next condition. The process repeats these steps until all conditions are not satisfied. The inference, then, goes to next phase and repeats the above process until all phases are tested.

In going to the next phase, the system can use either described algorithm or primitive command. A rule used in this knowledge base is a composition of primitive commands which are classified into seven groups as follows.

Loading Value Command is a group of commands used for loading and reading the specified value from dictionary, tables and IL. Operation Command is a group of commands used for 'set' operation. Moving Command is a group of commands used for moving the inference engine to the next rule or the next node. Cutting Command is a group of commands used for cutting the value of the variable or unwanted node. Changing Command is a group of commands used for changing the value of the node. Checking Command is a group of commands used for checking the type of node or the node location. Other Commands such as the command used for changing the structure of IL from a tree structure to a linear structure.

4. **<u>Results</u> <u>and</u> <u>Conclusion.</u>** This system is developed and tested with 50 sample sentences. There are 196 words in the dictionary. The study has indicated that the generated target language, Thai, is accurate and reliable in representing the source language. The followings are the output of the test run. Fig. 17 is the output of syntactic structure process of which input shown in fig. 16. Fig. 18 is the Thai words selection output and the translation output is shown in fig.19.

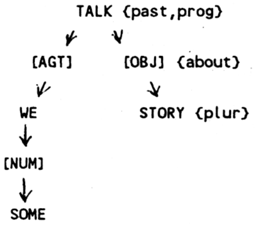


Fig. 16 The interlingua input.

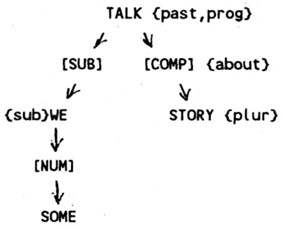


Fig. 17 The syntactic structure output.

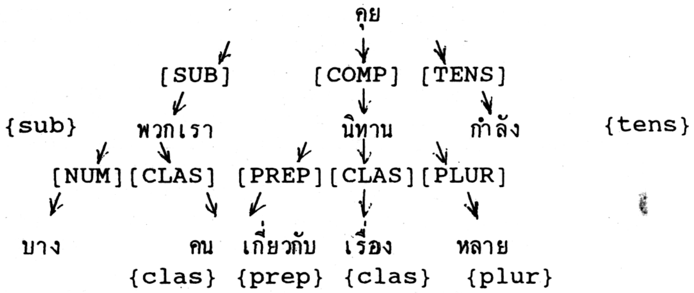


Fig. 18 The Thai words selection output.

พวกเรา บาง คน กำลัง คุย เกี่ยวกับ นิทาน หลาย เรื่อง

Fig. 19 The output of the system.

This study is only a prototype of an Interlingua Machine Translation system. The source language chosen is a page long story for children. The level of language is easy enough to be manageable in terms of language analysis and generation. The knowledge base of the system needs to be expanded to cope with the more complex structures of language. The primitive commands of the knowledge base must be revised to be more effective. However, the improvement of the system as mentioned must be done when we move our system from personal computer to engineering workstation computer.

References

Nirenburg, S. (1989) 'Knowledge-Based Machine Translation.' Machine Translation. Vol 4 1-24.

Nirenburg, S. (1987) **Machine Translation.** Cambridge University Press.

Nyberg, E, 3rd., McCardell, R., Gate, D., Nirenburg, S (1989) 'Generation.' Machine Translation Vol 4, 149-168.

Rich, E. (1983) **Artificial Intelligence.** McGraw-Hill.