Computational Intelligence, Volume 15, Number 2, 1999

ENGLISH-THAI TRANSLATION: INITIAL EXPERIMENTS WITH A MULTIPHASE TRANSLATION SYSTEM

Kanlaya Naruedomkul
Computer Science Department, University of Regina

Nick Cercone
Computer Science Department, University of Waterloo

Booncharoen Sirinaovakul
Computer Engineering Department, King Mongkut's University of Technology Thonburi, Bangkok

A multiphase machine translation approach, Generate and Repair Machine Translation (GRMT), is proposed. GRMT is designed to generate accurate translations that focus primarily on retaining the linguistic meaning of the source language sentence. GRMT presently incorporates a limited multilingual translation capability. The central idea behind the GRMT approach is to generate a translation candidate (TC) by quick and dirty machine translation (QDMT), then investigate the accuracy of that TC by translation candidate evaluation (TCE), and, if necessary, revise the translation in the repair and iterate (RI) phase. To demonstrate the GRMT approach, a translation system that translates from English to Thai has been developed. This paper presents the design characteristics and some experimental results of QDMT and also the initial design, some experiments, and proposed ideas behind TCE and RI.

Key words: machine translation, multilingual machine translation, direct approach, transfer approach, interlingual approach, HPSG.

1. INTRODUCTION

Accurate machine translation (MT) and multilingual machine translation (MMT) systems are increasingly in demand as transportation and trade between countries "shrink" our world and make good communications between population groups speaking different languages more important. However, existing MT systems are still far from ideal because of limitations of MT approaches. The three classic approaches: Direct, Transfer, and Interlingual each has its own disadvantages.

In the Direct approach, which is a word-to-word replacement strategy, the accuracy of the translation is rather limited because it takes into account only morphological information. The following examples illustrate this limitation.

	Source Language	Target Language
1.	open door	เปิดประตู (pэ`эd-open) (pratuu door)
2.	open fire	เริ่มยิง (rэ^эm-open) (jiη-fire)
3.	open arms	อ้าแขน (?âa-open) (khεˇεn- arm)
4.	open eyes	ลืมตา (lyym-open) (taa-eye)

Address correspondence to Kanlaya Naruedomkul, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; e-mail: kanlaya@cs.uregina.ca.

5. เปิดไฟ turn on the light
(pɘ`ɘd-turn on) (faj´-light)

6. ลืมบอก forget to tell
(lyym-forget) (b∂`∂g´-tell))

Example 1, the word "open" is translated as "เปิด- pɘ´ɘd" in Thai because it means "not shut." However, the word "open" in example 2 is translated as "เริ่ม- rɘ^ɘm" since the phrase "open fire" means "to start shooting." Nevertheless, the translation of the word "เปิด- pɘ`ɘd," example 5, into English is "turn on," not "open." The word "ลืม- lyym," example 6, cannot be translated as "open" either because it has a completely different meaning from the word "ลืม- lyym" in example 4. The translation of the word "ลืม- lyym" in example 6 must be "forget."

The meaning of a word or phrase depends, in part, on the context in which it is used. The relation between the basic meaning of a word and its circumstances of usage is very complex. This complexity is one reason why the mapping from one language to another is far from direct.

The Transfer approach, more linguistically motivated, transfers the meaning and structure of the source language (SL) to the target language (TL) via the internal representation of each by "transfer rules." The translation result is more accurate than the results produced by the Direct approach but the results are generally inadequate because the accuracy depends mainly on the transfer process. It is possible that the system may transfer unnecessary information that may not be needed to generate the TL; meanwhile the system might lose some information that is needed in the TL due to the different characteristics between languages and the complexity of the transfer process. This problem can be seen clearly from the following sentences:

1. It does not matter if you are born in a duck yard.
2. *Il n'importe pas si vous naissez dans un jardin du canard*.
3. He/it doesn't import if you are born in a garden of the duck.

Sentence 2 is the translation into French of sentence 1, and sentence 3 represents the translation of sentence 2 back into English.[1] Translation sentence 2 does not retain the meaning of the original sentence 1 properly. It is incorrect both in the words selected and in grammar. Translation sentence 3 is not the correct translation of sentence 2 either. Translation 3 should convey the same meaning as sentence 1 but it does not because of errors accumulated during the two translation processes. The Transfer approach does not appear appropriate for multilingual systems because it requires a set of transfer rules for each language pair.

The Interlingual (IL) approach generates the target language from the intermediate representation, which is totally independent of language pairs. The interlingual idea is the most attractive to the MMT system. However, to define neutral concepts for different languages is rather a chimera. Many interlingual systems have been developed: for example, ATLAS of Fujitsu (Uchida 1989), PIVOT of NEC (Muraki 1989), Rosetta of Phillips (Landsbergen 1987), KANT (Mitamura, Nyberg, and Carbonell 1991), and CICC interlingual (CICC 1995c). The CICC interlingual system was developed and implemented in the project "Research and development cooperation project on a Machine Translation System for Japan and its neighboring countries[2] 1987–1994" (CICC 1995d). This project sought to develop MMT that could translate Japanese, Chinese, Malaysian, Indonesian, and Thai. Some limitations and problems were encountered in developing the interlingual system for this project. For

[1] These translations were provided by a commercial MT system.

[2] China, Indonesia, Malaysia, and Thailand.

example, interlingua theory assumes that sentences in different languages which carry the same meaning must be represented by the same interlingua. However, in the CICC MMT project, the interlingual representations of these sentences are different (CICC 1995b) for several reasons (Cercone 1975; CICC 1995a; CICC 1995d; CICC 1995e; CICC 1995f):

- The scopes and concept classification of each language are different (e.g., the concept of culture, the concept of the supernatural world, and the concept of unit).
- The linking of concepts between languages cannot be handled by the existing system.
- A single concept in one language can be mapped into many concepts in another language.
- More than one concept can be mapped into only one single concept in another language.
- Some concepts do not exist in some languages.

The following examples illustrate that some concepts do not exist in some languages. A Thai salute of greeting or leave-taking is called "ไหว้." It is performed by placing the hands together at the chest or raising them toward the face. This kind of salute does not exist in most cultures. Similarly, a part of one Thai national costume is a piece of cloth wrapped around the chest and back. This piece of cloth is called "สไบ," but such apparel does not exist in other nations' costumes.

It is obvious that differences in languages and cultures result in different circumstances of language usage. Thus, it is very difficult to define a neutral concept. The interlingual approach, then, is still an ideal despite being considered the best approach.

The nonlinguistic MT approaches such as statistical strategies (Brown et al. 1992) and example-based MT (Jones 1996) and the hybrid approaches such as knowledge-based MT (Goodman and Nirenburg 1991) are interesting, but have not yet proved deployable for accurate large-scale MT and MMT systems.

In order to increase the accuracy of translation, avoid the difficulties in developing an IL system, and promote MMT, Naruedomkul and Cercone (1997) proposed Generate and Repair Machine Translation (GRMT). GRMT generates a translation candidate (TC) in the target language and compares the meaning of the translation candidate with the meaning of the corresponding source language. If there is no significant dissimilarity then that TC must be an appropriate translation, otherwise the TC is repaired. The meaning of the repaired TC is again compared with that of the source language. Comparison and repair processes are repeated until an accurate translation is achieved.

2. GENERATE AND REPAIR MACHINE TRANSLATION

GRMT (Figure 1) is designed to serve two purposes: to generate an accurate translation and to be amenable to multilingual translation. An accurate translation corresponds to a translation that retains the linguistic meaning of the SL. To achieve an accurate translation, GRMT performs the translation in three phases: The first phase, Quick and Dirty Machine Translation, generates the translation candidate for the source language. The accuracy of the generated TC is evaluated by analyzing both the TC and the SL in the second phase, Translation Candidate Evaluation (TCE). TCE compares the semantic information of the TC with that of the SL. If there is any dissimilarity, the TC is "repaired" in the third phase, Repair and Iterate (RI). The repaired TC is again analyzed and compared until there is no appreciable dissimilarity between the meaning of the SL and that of the TC. By performing TCE and RI processes, GRMT ensures an accurate translation.

GRMT treats the source and target languages separately and is aware of differences between languages. Therefore, if we group languages according to the various characteristics

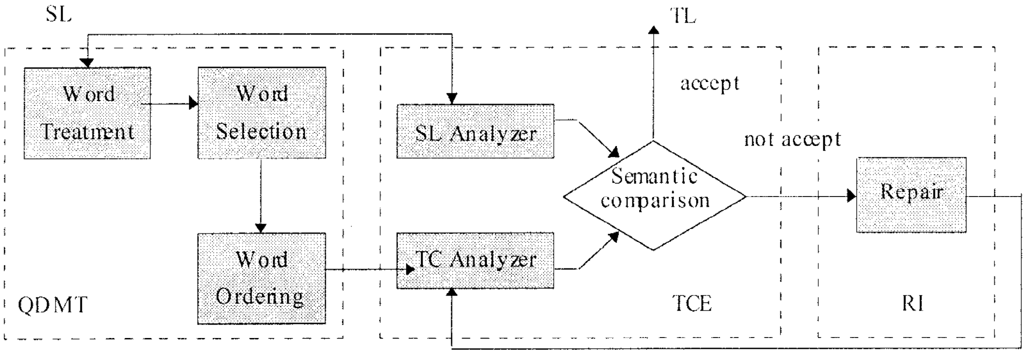


FIGURE 1. GRMT architecture.

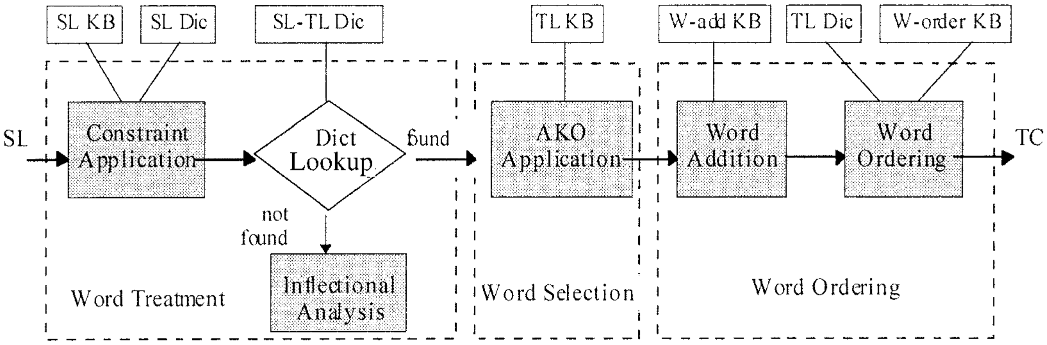


FIGURE 2. QDMT architecture.

they have in common (some examples are given in Section 3.1), then we can perform the translation between groups more simply by GRMT. For example: Group 1 consists of English, French, and Spanish, Group 2 consists of Chinese, Japanese, and Thai. To perform the translation between these two groups, the transfer approach requires six SL analyzers, six TL generations, and 18 sets of transfer rules; GRMT requires six SL analyzers, six TL analyzers, and two sets of constraint applications.

3. QUICK AND DIRTY MACHINE TRANSLATION

Quick and dirty machine translation generates a translation candidate by considering the difference between language pairs in terms of syntax and semantics without performing any sophisticated analysis. QDMT first considers target language words that correspond to all possible meanings of each source language word. The most appropriate TL word is selected by applying a semantic relationship between words and then the selected words are rearranged according to the grammar of the TL. As is shown in Figure 2, QDMT comprises three modules: word treatment, word selection, and word ordering.

Word treatment: There are two steps performed by this module: SL constraint application and dictionary lookup. SL constraints (Naruedomkul and Cercone 1997) are applied when characteristics of the SL differ from those of the TL (e.g., auxiliary verb, passive voice).

These constraints are used to narrow the scope of possible TL words that correspond to each SL word.

After applying the constraints, each word is used as a key word to search for the corresponding word in the TL. If the key word used can be found in the bilingual dictionary, all possible corresponding TL words are attached to that SL word. If the key word cannot be found, inflectional analysis is performed before searching anew. Inflectional analysis provides information about tense, plurality, present participle, and comparison for such input words. It also indicates the part of speech of the word—for example, verb for tense and present participle, noun for plural, and adjective for comparison. This information is useful in the word selection step.

Word selection: QDMT selects the most appropriate word for each input word by considering the semantic relationship between the close proximity words. Words in close proximity are considered to have stronger connections. The semantic relationship indicates which word can occur with which word. We have designed the semantic relationship (illustrated in Section 3.3) by using the “A Kind of” (AKO) slot (Tantisawetrat and Sirinaovakul 1991) from the CICC MMT project; AKO is a classification of words based on meaning. We augmented their usage to include information used for word selection. Our detailed algorithm for word selection is presented in Naruedomkul and Cercone (1997).

Word ordering: Two steps are performed by this module: word addition and word ordering. The differences between languages is reconsidered at this point to complete the target language. Some words that are necessary in the TL, not only to retain the meaning of the source language but also to make them grammatically correct, are added into the string before the ordering can be performed. For example, past tense is shown by verb inflection in English, but in Chinese, Japanese, and Thai no inflection results from tense. Past tense can be expressed by using a modifying verb (e.g., “le” in Chinese, “itta” in Japanese, and “lε´εw” in Thai.[3])

The selected words are rearranged in a grammatical order according to the ordering rule without performing any analysis. QDMT ordering rules are generated from a number of TL examples consistent with the TL grammar. The ordering is considered based on the subcategory information of words. The subcategory is a more narrowly defined function of the word. The ordered string results in a translation candidate. The results of word ordering are illustrated in Section 3.1.4.

3.1. Experimental Results

Our experimental results were generated by performing translation from English into Thai. Therefore, in this section, English is regarded as the source language and Thai is regarded as the target language. Examples of SL constraints, dictionaries, the semantic relationship, and ordering rules that have been developed for QDMT are discussed in this and subsequent sections.

SL Constraints. The SL constraint is required in the word treatment module. Some characteristics of English that are different from Thai include:

1. Auxiliary Verb Constraint: Auxiliary verbs in English are needed in many cases, as in front of an adjective or the negative “not,” but for the same expression they are not used

[3] Phonetic transcriptions of Chinese, Japanese and Thai, respectively.

in Thai, as shown in the following examples:

be + adjective → adjective
I am glad. → ฉัน ดีใจ
(chân- I) (diicaj- glad)

do + not → not
He does not eat. → เขา ไม่ กิน
(khaw- he) (mâj - not) (kin- eat)

2. Present Continuous and Present Perfect (Continuous) Form Constraint: In English, the inflection "ing" form of a verb is needed after an auxiliary verb to describe an action that is going on at the moment of speaking. However, the Thai language does not have this inflection; therefore, the word "กำลัง -kamlaη" is used to describe the same action without changing the verb form. For example:

be + V ing → กำลัง + V
I am swimming. → ฉัน กำลัง ว่ายน้ำ
(chân-I) (kamlaη-ing) ('wâaj náam' -swim)
(Exception: interesting, . . .)

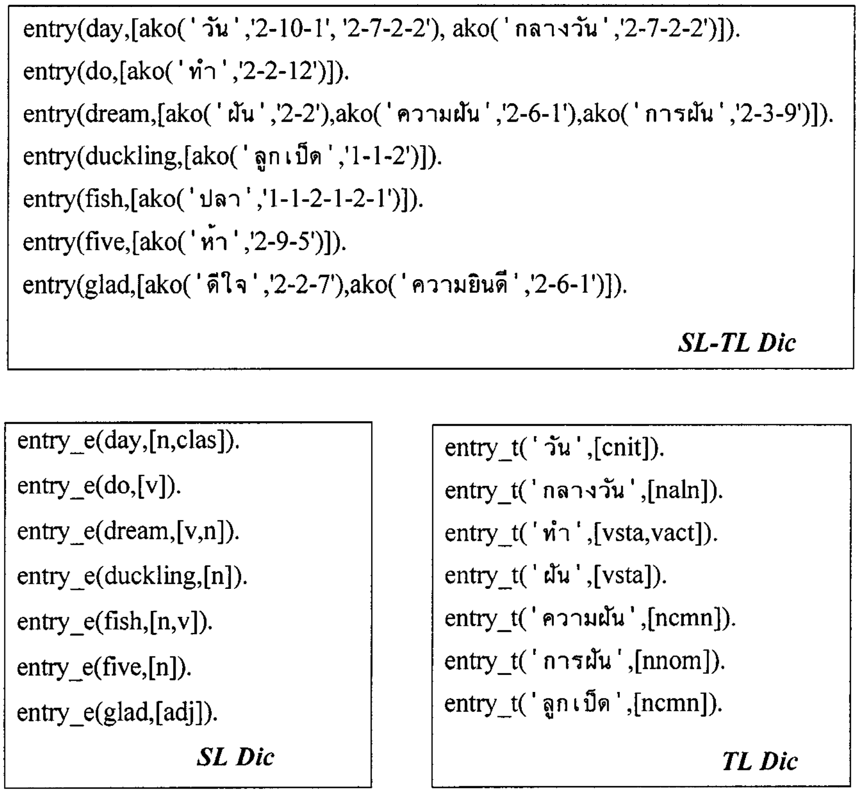
3. Passive Voice Constraint: Another constraint is the passive voice, which is used in English but it is rarely used in Thai. Passive voice in Thai is used mainly when discussing an unpleasant situation; otherwise it does not carry any meaning.

The book was taken by him. → หนังสือ ถูก เขา เอา ไป
(naηsyˇy -book) (thùug -passive) (khaw -him)
(?aw -take) (paj - modifying)
เขา เอา หนังสือ ไป
(khaw -him) (?aw -take) (naηsyˇy -book)
(paj - modifying)

The word "ถูก" denotes the passive voice in Thai. The first of these two Thai sentences, which is passive, is not generally used in Thai; the second Thai sentence carries the same meaning but it is in active voice.

3.2. Dictionaries

Three dictionaries are needed in QDMT: the SL dictionary, the SL-TL dictionary, and the TL dictionary. Entries in the SL and TL dictionaries can be single words and some inflected and derived forms that cannot easily be handled by morphological rules. Compound words are also included. Each entry in the SL dictionary contains a word form and the word category information needed for the inflectional analysis step. Figure 3 shows examples of the three dictionaries. The Thai dictionary entry contains the word form and word subcategory used in the ordering step. The SL-TL dictionary contains the English entry and all corresponding Thai words and AKO numbers of each Thai word; for example, the word "dream" in English has three corresponding Thai words, which express differences in meaning and usage. All Thai words that correspond to each English entry are ordered based on the frequency of usage



```
entry(day,[ako('วัน','2-10-1', '2-7-2-2'), ako('กลางวัน','2-7-2-2')]).
entry(do,[ako('ทำ','2-2-12')]).
entry(dream,[ako('ฝัน','2-2'),ako('ความฝัน','2-6-1'),ako('การฝัน','2-3-9')]).
entry(duckling,[ako('ลูกเป็ด','1-1-2')]).
entry(fish,[ako('ปลา','1-1-2-1-2-1')]).
entry(five,[ako('ห้า','2-9-5')]).
entry(glad,[ako('ดีใจ','2-2-7'),ako('ความยินดี','2-6-1')]).
```

SL-TL Dic

```
entry_e(day,[n,clas]).
entry_e(do,[v]).
entry_e(dream,[v,n]).
entry_e(duckling,[n]).
entry_e(fish,[n,v]).
entry_e(five,[n]).
entry_e(glad,[adj]).
```

SL Dic

```
entry_t('วัน',[cnit]).
entry_t('กลางวัน',[naln]).
entry_t('ทำ',[vsta,vact]).
entry_t('ฝัน',[vsta]).
entry_t('ความฝัน',[ncmn]).
entry_t('การฝัน',[nnom]).
entry_t('ลูกเป็ด',[ncmn]).
```

TL Dic

FIGURE 3. Examples of dictionaries.

(in real life). The first meaning is selected once the constraint and AKO fail. However, all dictionaries will be modified to some extent to serve the second and the third phases of the GRMT.

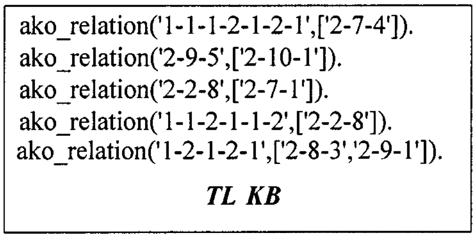
The developed dictionaries are compiled from a Thai-English dictionary[4] that was developed for the CICC MMT project, the On-line MT Dictionary (Thai/English beta test),[5] the New Model English-Thai Dictionary (Sethaputra 1977), and the Thai-English Student's Dictionary (Hass 1964). Our compilation is based on words that are currently in use, since some dictionary entries are archaic. Another feature of our compilation is that multiple entries of common meaning have been simplified.

3.3. Semantic Relationship

A semantic relationship (illustrated in Figure 4) contains AKO numbers of words that can occur in the same expression. The word with AKO number in the first argument can occur with the word that has an AKO number shown in the second argument. For example, in the phrase "five days" the word "day" can be translated as "กลางวัน -klaaηwan" or "วัน -wan" in Thai. The AKO values of these two words are different because they have

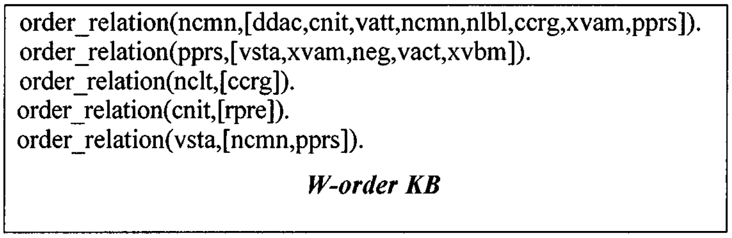
[4] This dictionary was provided by the Linguistics and Knowledge Science Laboratory, NECTEC, Bangkok, Thailand.

[5] This dictionary was provided by Virach Sornlertlamvanich, Tokyo Institute of Technology, Tokyo, Japan.



```
ako_relation('1-1-1-2-1-2-1',['2-7-4']).
ako_relation('2-9-5',['2-10-1']).
ako_relation('2-2-8',['2-7-1']).
ako_relation('1-1-2-1-1-2',['2-2-8']).
ako_relation('1-2-1-2-1',['2-8-3','2-9-1']).

TL KB
```

FIGURE 4. Examples of a semantic relationship.



```
order_relation(ncmn,[ddac,cnit,vatt,ncmn,nlbl,ccrg,xvam,pprs]).
order_relation(pprs,[vsta,xvam,neg,vact,xvbm]).
order_relation(nclt,[ccrg]).
order_relation(cnit,[rpre]).
order_relation(vsta,[ncmn,pprs]).

W-order KB
```

FIGURE 5. Examples of ordering rules.

different meanings: "กลางวัน" means "the time between sunrise and sunset," with AKO value 2-7-2-2; "วัน" means "a period of 24 hours" with AKO value 2-7-2-2 and it also can be used as a classifier in Thai with AKO value 2-10-1. The word "five" is translated to "ห้า -hâa" with AKO value 2-9-5 and the semantic relationship shows that this word can occur with the word that has AKO value 2-10-1 (line 2 of Figure 4). So the appropriate word "วัน" is selected as indicated by the AKO value of "five."

3.4. Ordering Rules

QDMT rearranges the selected words in a grammatical order by considering subcategory information of words. The subcategory is a more narrowly defined function of the word. For example, the determiners are classified into nine subcategories: "ddan," "ddac," "diac," etc. (Tantisawetrat and Sirinaovakul 1991). The determiners นี่ (nîi-this), นั่น (nân-that), and โน่น (nôon-those) are classified as "ddan." The determiners นี้ (níi-this), นั้น (nán-that), and โน้น (nóon-those) are classified as "ddac." These determiners can be structured as follows:

1. "ncmn" + "ddan"
2. "ncmn" + {classifier} + "ddac"

where "ncmn" is a subcategory of noun; for example, โต๊ะ (tó?-table), หนังสือ (naηsyˇy-book), ลูกเป็ด (lûugpèd-duckling).

In the first case there would not be a classifier[6] present; see, for example, sentence 1 below. In the second case a classifier may be present but it is optional (Punmetha 1984); see,

[6] "Classifier" indicates the unit of a countable noun (see Section 4.1).

TABLE 1. QDMT Steps Applied—Example 1.

Example 1. Algebraic symbols are used when you do not know what you are talking about.

TC: สัญญลักษณ์ ทางพีชคณิต ถูก ใช้ เมื่อ คุณ ไม่ รู้ ว่า คุณ กำลัง พูด เกี่ยวกับ อะไร

CT: สัญญลักษณ์ ทางพีชคณิต ถูก ใช้ เมื่อ คุณ ไม่ รู้ ว่า คุณ กำลัง พูด เกี่ยวกับ อะไร
(sanjalág-symbol) (thaaη phichakhaníd-algebraic) (thùug-passive) (cháj-use) (myˆa-when) (khun-you) (mâj-not) (rúu-know) (wâa-connective) (khun-you) (kamlaη-ing) (phûud-talk) (kiàw kàb-about) (araj-what)

Input	Constraint application	Dict. lookup and inflec. analysis	Word selection	Selected word
Algebraic symbols	Algebraic symbols	ทางพีชคณิต สัญญลักษณ์ เครื่องหมาย	ทางพีชคณิต สัญญลักษณ์ เครื่องหมาย	ทางพีชคณิต สัญญลักษณ์
are	passive	ถูก	ถูก	ถูก
use	use	ใช้ ประโยชน์	ใช้	ใช้
when	when	เมื่อ, เมื่อไหร่, ขณะที่	เมื่อ เมื่อไหร่ ขณะที่	เมื่อ,
you	you	คุณ	คุณ	คุณ
do	do	—	—	—
not	not	ไม่	ไม่	ไม่
know	know	รู้ รู้จัก	รู้	รู้
what	what	อะไร	อะไร	อะไร
you	you	คุณ	คุณ	คุณ
are	ing	กำลัง	กำลัง	กำลัง
talking	talk	พูด	พูด	พูด
about	about	ประมาณ, เกี่ยวกับ รอบๆ	เกี่ยวกับ,	เกี่ยวกับ

for example, sentence 2 below. Both sentences 1 and 2 convey the meaning “This table is bigger than that table.”

1. โต๊ะ นี่ ใหญ่ กว่า โต๊ะ นั่น
(tó?-table) (nîi-this) (jàj-big) (kwàa- than) (tó?-table) (nán-that).
2. โต๊ะ {ตัว} นี้ ใหญ่ กว่า โต๊ะ {ตัว} นั้น
(tó?-table) (tua-clas) (níi-this) (jàj -big) (kwàa - than) (tó?-table) (tua-clas) (nán-that).

Figure 5 shows examples of ordering rules. The second argument of a rule is a list of subcategories of words following any word that has the subcategory shown in the first argument. For example, the translation into Thai of “long leg” is “ขา ยาว- khaˇa jaaw,” which is literally transposed as “leg long.” The ordering of Thai is different from that of

English because the ordering rule indicates that the subcatgory vatt[7] of the word “ยาว (jaaw-long)” must follow the word “ขา (khaˇa-leg),” which has the subcategory “ncmn.”

3.5. Examples of QDMT

To illustrate the performance of QDMT, an initial version of the translation from English to Thai has been developed and run under SICStus Prolog 2.1, on a SUN workstation (because the existing Thai keyboard map works properly only on SUN workstations). The QDMT was tested to generate the translation candidate for a number of sentences by using the developed dictionary, which contains 152 English words and 348 Thai words. Some examples are shown in Tables 1 through 4. The SL symbols are shown in the first column, the selected TL words are shown in the last column. The second and the third columns show the results of constraint application and dictionary lookup steps. The words that were selected by semantic relationship are shown in the fourth column. The generated TC is presented and compared with the correct translation (CT) of the SL in each example.

In Example 1, three constraints were applied: “are used” triggers the “passive voice” constraint, “do not” triggers the “negative” constraint, and “are talking” triggers the “present continuous” constraint. Each word in the second column is used as a key word to search for the corresponding words in Thai. The word “symbols” is analyzed in terms of “plurality” before it can be found in a bilingual dictionary (Naruedomkul and Cercone 1997). All possible meanings of each input word are shown in the third column. Some of the input words have more than one meaning (“symbol,” “use,” “when,” “know,” “about”). The appropriate meaning of “use,” “know,” and “about” can be selected by considering the semantic relationship between words, and the choice for each of these is shown in the fourth column. However, the appropriate words for “symbol” and “when” cannot be selected in the same manner because all possible meanings of each word have the same AKO number. Therefore, the first meaning that appears in the bilingual dictionary of each is selected. All selected words are shown in the last column. Before performing the ordering step, the word “ว่า - wâa” is added to combine clauses. The word “ว่า” is a translation of the word “that,” which is omitted in this sentence; however, omitting it in the Thai the translation is grammatically incorrect.

In Examples 1 and 2, QDMT formed the correct translation for the input sentence without performing any correction. In Examples 3 and 4, the word selection is correct but the ordering of some words is not appropriate because the ordering is not yet complete.

In Example 2, the word “ห้อง” is added as a classifier. In Thai, a classifier is needed in front of the indefinite determiner “หนึ่ง.” Another classifier “ตัว” is added in Example 3. Different nouns relate to different classifiers (see Section 4.1 below for details).

In generating the TC for Example 4, once all words are selected, the words “ที่. . . จะ (thîi. . .cà?)” must be added to clarify tense. These additions are necessary because the preposition “before” shows that the “fishing” action would be taken after John took the worms. The generated TC is grammatically correct. However, this sentence is not the way it is spoken in Thai. The correct translation is ordered in the following way: “John took the worms before he went fishing.”

Table 5 shows some TCs that were generated by QDMT. The CTs are shown in italics. Some TCs do not require any repair; some TCs do. Example 2 presented in Table 2, and sentences 1 and 2 from Table 5 show that QDMT can select the appropriate Thai words for the different meanings of the word “rent” in each sentence. In Example 2, “rent” means “to

[7] “vatt” is a subcategory of verb.

TABLE 2. QDMT Steps Applied—Example 2.

Example 2. I rent a room from Mrs. Jones.

TC: ฉัน เช่า ห้อง ห้อง หนึ่ง จาก นาง โจนส์

CT: ฉัน เช่า ห้อง ห้อง หนึ่ง จาก นาง โจนส์
(chˇan-I) (chˇaw-rent) (h∂ˆη-room) (h∂ˆη-clas) (nyη-a) (ca`ag-from) (naaη-Mrs.) (jones -Jones)

Input	Constraint application	Dict. lookup and inflec. analysis	Word selection	Selected word
I	I		ฉัน	ฉัน
rent	rent	เช่า, ค่าเช่า, ให้เช่า, รอยขาด	เช่า	เช่า
a	a	หนึ่ง	หนึ่ง	หนึ่ง
room	room	ห้อง, ที่ว่าง, ช่องว่าง	ห้อง	ห้อง
from	from	จาก, ตั้งแต่, ออกจาก	จาก	จาก
Mrs.	Mrs.	นาง	นาง	นาง
Jones	Jones	โจนส์	โจนส์	โจนส์

TABLE 3. QDMT Steps Applied—Example 3.

Example 3. The ugly duckling hides his head under his wing.

TC: ลูกเป็ด ตัว นั้น ขี้เหร่ ซ่อน หัว ของเขา ใต้ ปีก ของเขา

CT: ลูกเป็ด ขี้เหร่ ตัว นั้น ซ่อน หัว ของเขา ใต้ ปีก ของเขา
(lûugpèd-duckling) (khîirèe-ugly) (tua-classifier) (nán-the) (s∂ˆ∂n-hide) (huˇa -head) (kh∂∂η khaw-his) (tâaj-under) (pìig -wing) (kh∂∂η khaw-his)

Input	Constraint application	Dict. lookup and inflec. analysis	Word selection	Selected word
The	The	นั้น	นั้น	นั้น
ugly	ugly	ขี้เหร่, น่าเกลียด	ขี้เหร่	ขี้เหร่
duckling	duckling	ลูกเป็ด	ลูกเป็ด	ลูกเป็ด
hides	hide	ซ่อน, ที่ซ่อน, การซ่อน, หนังสัตว์, การส่องสัตว์	ซ่อน	ซ่อน
his	his	ของเขา	ของเขา	ของเขา
head	head	หัว	หัว	หัว
under	under	ใต้, ภายใต้	ใต้	ใต้,
his	his	ของเขา	ของเขา	ของเขา
wing	wing	ปีก, ความปกป้อง, บิน, กองบิน	ปีก	ปีก

take and hold under and agreement to pay rent,” which corresponds to the word “เช่า-chˇaw” in Thai. In sentence 1, Table 5, “rent” means “the amount of money paid or due for the use of another’s property,” which is translated as “ค่าเช่า-khâachˇaw;” in sentence 2 it means “a tear in cloth” and is translated into Thai as “รอยขาด-r∂∂jkhàad.” However, in sentences 2,

TABLE 4. QDMT Steps Applied—Example 4.

Example 4. Before he went fishing John took the worms.

TC: ก่อน ที่ เขา จะ ไป ตกปลา จอห์น เอา ไส้เดือน

CT: จอห์น เอา ไส้เดือน ก่อน ที่ เขา จะ ไป ตกปลา
(c∂∂n-John) (?aw-take) (sâjdyan-worm) (k∂`∂n-before) (thîi-modifying) (khaw-he) (cà?-modifying) (paj-go) (tògplaa-fishing)

Input	Constraint application	Dict. lookup and inflec. analysis	Word selection	Selected word
Before	Before	ก่อน	ก่อน	ก่อน
he	he	เขา	เขา	เขา
went	went	ไป	ไป	ไป
fishing	fishing	ตกปลา	ตกปลา	ตกปลา
John	John	จอห์น	จอห์น	จอห์น
took	took	เอา, หยิบ, จับ, ลาก, พา	เอา	เอา
the	the	นั้น	นั้น	นั้น
worms	worm	ไส้เดือน, คนใจอ่อน	ไส้เดือน	ไส้เดือน

4, and 5 of Table 5, some words are not put in the right place by QDMT because we have not yet designed a complete set of ordering rules.

4. TRANSLATION CANDIDATE EVALUATION

The generated TC, which is an output of the previous step, QDMT, is analyzed to determine whether the TC retains the meaning of the SL. Our initial idea for TCE is to perform the evaluation by analyzing both the TC and the SL in terms of syntax and semantics in parallel (Figure 6), then compare the semantic results only (there are syntactic level differences between languages). If the semantic results are the same, that TC is an acceptable translation. If not, any part of the TC that causes its semantics to differ from that of SL is repaired in the next phase, RI.

The meaning of any phrase can be represented by the following features: semantic mode, situation index, and restriction (Sag and Wasow 1997). The semantic mode can be classified as follows: as proposition for the noninverted sentence, as question for the inverted sentence, as directive for the imperative phrase, and as reference for the noun phrase. A situation index is an index that corresponds to the situation referenced. Restriction specifies a list of conditions that the situation must satisfy, such as relation, instance, possessor, possessed.

Figure 7 illustrates the expected semantic representation of the source language of Example 5. Since the corresponding generated target candidate is an appropriate translation, we therefore expect the same representation. This semantic representation means the SL represents a proposition in which a situation s_1 satisfies the conditions that j is a car, j belongs to a female i, and i likes j.

Figure 8 illustrates the semantic representation of the source language and target language of Example 6. This semantic representation means the SL (or the TL) depicts a proposition

TABLE 5. Some Translation Candidates Generated by QDMT.

1. They will pay more rent.
   พวกเขา จะ จ่าย ค่าเช่า เพิ่ม
   (phûagkhaˇw-they) (cà?-will) (càaj-pay) (khâachâw-rent) (ph∂ˆ∂m-more)
   พวกเขา จะ จ่าย ค่าเช่า เพิ่ม

2. There are several rents in these trousers.
   มี รอยขาด หลาย รอย ใน กางเกง นี้ ตัว
   (mii-there are) (r∂∂j khàad-rent) (laˇaj-several) (r∂∂j-clas) (naj-in) (kaaη keeη-trousers) (níi-these) (tua-clas)
   มี รอยขาด หลาย รอย ใน กางเกง ตัว นี้

3. The beautiful Egyptian talked about her dream.
   คนอียิปต์ คน สวย นั้น พูดเกี่ยวกับ ความฝัน ของเธอ
   (khon ?ii´jìb-Egyptian) (khon -clas) (suˇaj-beautiful) (nán-the) (phûud-talk) (kiàw kàb-about) (khwaam faˇn-dream) (kh∂∂η thɔɔ-her)
   คนอียิปต์ คน สวย นั้น พูดเกี่ยวกับ ความฝัน ของเธอ

4. Mrs. Jones gives the duckling happiness.
   นาง โจนส์ ให้ กับ ลูกเป็ด นั้น ตัว ความสุข
   (naaη-Mrs.)(jone-Jone)(háj-give)(kab-prep)(lûugpèd-duckling)(nán-the)(tua-clas)(khwaamsùg-happiness)
   นาง โจน ให้ ความสุข กับ ลูกเป็ด ตัว นั้น

5. Here the stork marched about on his long red legs.
   ที่นี่ นกกระสา นั้น เดินแถว รอบๆ ที่ ขา ยาว ของเขา สีแดง
   (thîinîi-here) (nógkrasaˇa-stork) (nán-the) (d∂∂nth∂∂η-march) (r∂ˆ∂b r∂ˆ∂b-about) (thîi-on) (khaˇa-leg) (jaaw-long) (kh∂ˇ∂khaˇw-his) (siˇidεεη-red)
   ที่นี่ นกกระสา นั้น เดินแถว รอบๆ ด้วย ขา ยาว สีแดง ของเขา

6. Kim gave Sandy a book.
   คิม ให้ หนังสือ กับ แซนดี้
   (kim-kim) (háj-give) (naηsyˇy-book) (kab-prep) (sandy-sandy)
   คิม ให้ หนังสือ กับ แซนดี้

in which a situation s_1 satisfies the conditions that k is a wing, j is a head, j is hidden under k, k and j belong to a male duckling i, and i is ugly.

We modify the semantic features based on our HPSG grammar formalism, intending to make them suitable to represent the meaning of any phrase in any language.

4.1. HPSG augmentation for English and Thai grammars

We have developed an HPSG analyzer for English based on six schemas: the subject-head schema, the head-complement schema, the specifier-head schema, the head-subject-

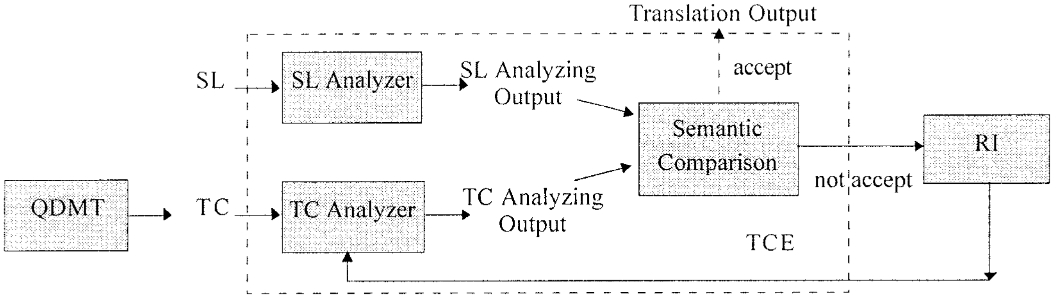


FIGURE 6. TCE architecture.

Example 5. She likes her car.

TC: เธอ ชอบ รถยนต์ ของเธอ
(thɘɘ-she) (ch∂ˆ∂b-like) (ródjon-car) (kh∂∂η thɘɘ-her)

$$
\begin{bmatrix}
\text{MODE} & \text{prop} \\
\text{INDEX} & s_1 \\
\text{RESTR} & \left\langle
\begin{bmatrix} \text{RELN} & \text{like} \\ \text{SIT} & s_1 \\ \text{LIKES} & i \\ \text{LIKED} & j \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{female} \\ \text{SIT} & s_2 \\ \text{INSTANCE} & i \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{female-poss} \\ \text{SIT} & s_3 \\ \text{POSSESSOR} & i \\ \text{POSSESSED} & j \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{car} \\ \text{SIT} & s_4 \\ \text{INSTANCE} & j \end{bmatrix}
\right\rangle
\end{bmatrix}
$$

FIGURE 7. Semantic representation of SL and TL—Example 5.

Example 6. The ugly duckling hides his head under his wing.

TC: ลูกเป็ด ขี้เหร่ นั้น ซ่อน หัว ของเขา ใต้ ปีก ของเขา
(lûugpèd-duckling) (khîirèe-ugly) (nán-the) (s∂ˆ∂n´-hide) (huˇa´ -head) (kh∂∂η´ khaw-his) (tâaj´-under) (pìig´ -wing) (kh∂∂η´ khaw-his)

$$
\begin{bmatrix}
\text{MODE} & \text{prop} \\
\text{INDEX} & s_1 \\
\text{RESTR} & \left\langle
\begin{array}{l}
\begin{bmatrix} \text{RELN} & \text{hide} \\ \text{SIT} & s_1 \\ \text{HIDES} & i \\ \text{HID} & j \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{ugly} \\ \text{SIT} & s_2 \\ \text{INSTANCE} & i \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{duckling} \\ \text{SIT} & s_3 \\ \text{INSTANCE} & i \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{male-poss} \\ \text{SIT} & s_4 \\ \text{POSSESSOR} & i \\ \text{POSSESSED} & j \end{bmatrix} \\
\begin{bmatrix} \text{RELN} & \text{head} \\ \text{SIT} & s_5 \\ \text{INSTANCE} & j \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{under} \\ \text{SIT} & s_6 \\ \text{LOWER} & j \\ \text{HIGHER} & k \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{male-poss} \\ \text{SIT} & s_7 \\ \text{POSSESSOR} & i \\ \text{POSSESSED} & k \end{bmatrix}
\begin{bmatrix} \text{RELN} & \text{wing} \\ \text{SIT} & s_8 \\ \text{INSTANCE} & k \end{bmatrix}
\end{array}
\right\rangle
\end{bmatrix}
$$

FIGURE 8. Semantic representation of SL and TL—Example 6.

complement schema, the adjunct-head schema, and the filler-head schema, which were provided by Matheson (1996).

The typical sentence structure of English and Thai is basically the same, with subject, verb, and object in that order; for example:

(1) เขา (2) ปล่อย (3) ปลา (1) He (2) released a (3) fish.
(khaˇw-he) (pl∂`j-release) (plaa-fish)

There are syntactic level differences between English and Thai; for example, in Thai the head usually comes before the attribute, as in the following example:

(1) หนังสือ (2) สีแดง The (2) red (1) book.
(naηsyˇy-book) (siˇidεεη-red)

Therefore, the head-adjunct schema was introduced to handle this structure.

Another feature of Thai that is different from English is the "classifier." The classifier indicates the unit of a countable noun. The classifier plays an important role in noun constructions that express a quantity or modify a noun. From our studies, the classifier can be categorized into six groups: type classifier, group classifier, feature classifier, measurement classifier, state classifier, and frequency classifier. There are more than 3,000 different classifiers in Thai. For example:

ปลา 2 ตัว two fish
(plaa-fish) (s∂ˇ∂η-two) (tua-clas)
หนังสือ 2 เล่ม two books
(naηsyˇy-book) (s∂ˇ∂η-two) (lêm-clas)
ปลา 1 ฝูง a group of fish
(plaa-fish)
หนังสือ 1 กอง a number of books.
(naηsyˇy-book)

The structure of a classifier can be:

1. noun + number or quantifier + classifier; for example,

หนังสือ 2 เล่ม two books
(naηsyˇy-book) (s∂ˇ∂η-two) (lêm-clas)
หนังสือ ทุก เล่ม every book.
(naηsyˇy-book) (thúg-every) (lêm-clas)

2. noun + (classifier) + adjective or specifier:

หนังสือ (เล่ม) สีแดง red book.
(naηsyˇy-book) (lêm-clas) (siˇidεεη-red)
หนังสือ เล่ม นั้น that book
(naηsyˇy-book) (lêm-clas) (nán-that)

To handle these structures, the head-numVquant-clas schema and the head-clas-adjVdet schema were introduced.

Also the structure of the lexicon was designed in order to serve these schemas. Each noun requires an appropriate classifier; an inappropriate classifier may not convey the meaning

```
lûugpèd
synsem  loc  cat   head      noun
                             mod     none
                             ako     112112
                   subj      [ ]
                   comps     [ ]
                   marking   unmarked
                   prev      [ ]
             cont  index     per     third
                             num     sg
                   restr     elt   nucleus   lûugpèd
                                             instance  per   third
                                                       num   sg
                                   quants    [ ]
                             elts  e_set
```

FIGURE 9. The lexical “ลูกเป็ด (lûugpèd-duckling).”

```
tua
synsem  loc  cat   head      clas
                             clas_ako    {112112, 115}
                   subj      [ ]
                   comps     [ ]
                   marking   unmarked
                   prev      [@ detp(Cont)]
             cont  clas      type
                   restind   Cont
```

FIGURE 10. The lexical “ตัว (tua-classifier).”

that the speaker intends to or may not convey any meaning. To select the proper classifier for a noun, the AKO number is used. Nouns that have the same AKO number relate to the same classifier; for example, “animal,” which has the AKO number 1-1-2, requires the classifier “ตัว-tua.” “Person,” with the AKO number 1-1-1-1, requires the classifier

TABLE 6. Subgroup of “Vehicle.”

AKO number	Examples	Classifier
1-2-5-1-1	รถยนต์ (ródjon-automobile) รถม้า (ródmáa-horse carriage) รถลาก (ródlâag-rickshaw)	คัน
1-2-5-1-2	เรือใบ (ryabaj-sailboat) เรือหางยาว (ryahaˇaηjaaw-“long-tailed boat”)	ลำ
1-2-5-1-3	รถไฟ (ródfaj-train)	ขบวน

“คน-khon.” Moreover, one classifier may fit into more than one set of AKO numbers; for example, the classifier “ตัว-tua” is compatible with “animal” (1-1-2-1-2-1) and also with “microorganism” (1-1-5). Therefore, to restrict the relationship between nouns and their classifiers, the features “AKO,” “clas_AKO,” and “prev” are introduced as shown in Figures 9 and 10. The feature “AKO” contains the AKO number of the object “lûugpèd-duckling”; Figure 9 shows the structure of the lexical entry for “lûugpèd-duckling” which has AKO number 112112. In Figure 10, the lexical entry “tua,” a classifier, is compatible with the set of nouns that have an AKO number specified by the feature “clas_AKO.” Also in Figure 10, “prev” is a list of words that can come before the object “tua.”

Because we have adopted the AKO approach developed in the CICC MMT project (Tantisawetrat and Sirinaovakul 1991), we require some additional analysis to modify this approach to account for a more comprehensive classifier analysis and treatment. For example, if we classify “vehicle” into three subgroups by mode of transportation as shown in Table 6, then the classifiers “คัน-khan,” “ลำ-lam,” and “ขบวน-khabuan” are compatible with each subgroup 1-2-5-1-1, 1-2-5-1-2, and 1-2-5-1-3, respectively.

4.2. Examples of TCE

To demonstrate the idea of TCE, the initial version of the Thai HPSG grammar has been implemented on the Attribute Logic Engine (ALE) (Carpenter and Penn 1994). The Thai analyzer has been run in parallel with the English HPSG analyzer. Figure 11 shows the results of analyzing the TC (หนังสือ 3 เล่ม) and SL (Three books) of Example 7. If a different TC is generated for Example 7, for example, if QDMT selects the classifier “ตัว (tua-clas)” for “หนังสือ (naηsyˇy-book),” resulting in the following TC:

หนังสือ 3 ตัว
(naηsyˇy-book) (saˇam-three) (tua-clas),

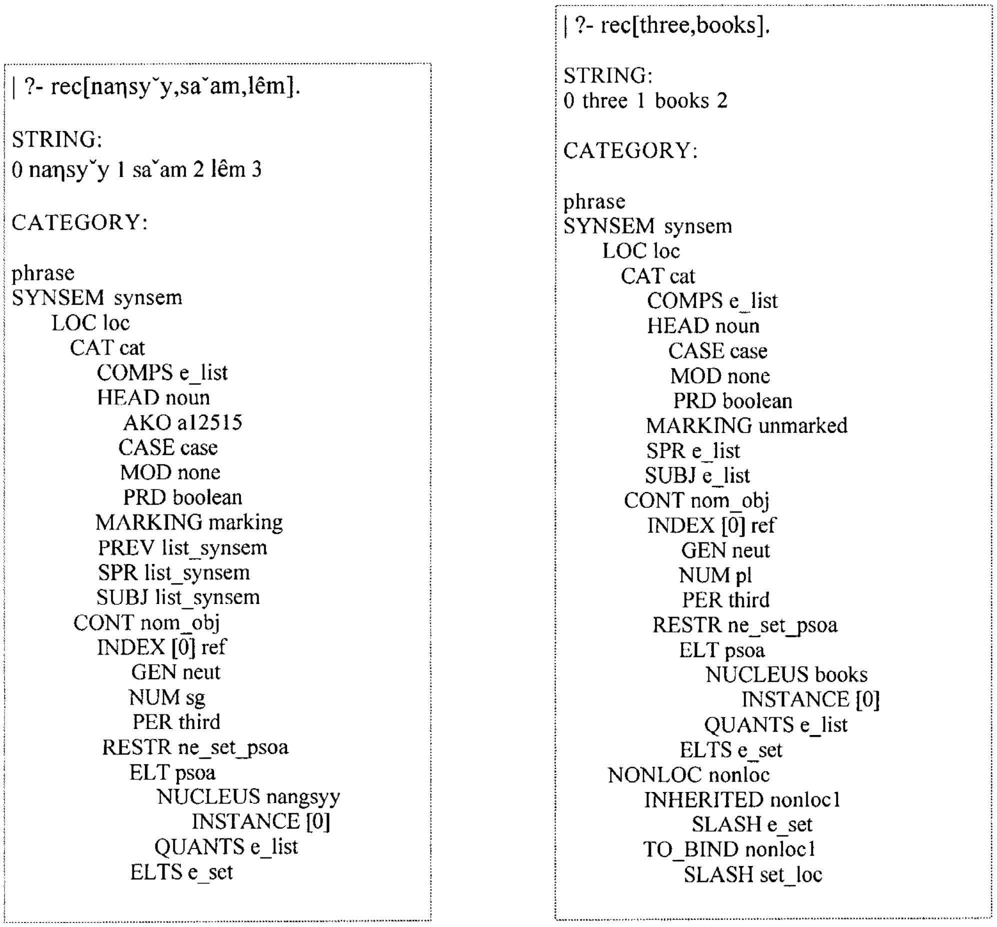
then Figure 12 shows that QDMT’s choice is unacceptable because the classifier “ตัว (tua-clas)” is not compatible and the TC fails to parse correctly.

Figure 13 shows the results of analyzing the translation pair (SL and TC) of Example 8. The semantic information of both parses (in the dashed boxes) shows that the “giver” is a masculine—(Kim), the “given” is a feminine—(Sandy), and the “gift” is classified as a neutral object—(book). However, this semantic representation does not provide a fine-grained representation of the meaning of the expression; for example, in the dashed box of part (a) the feature “gift” could be satisfied by a variety of nouns, etc. We would like to restrict the range of nouns that this feature describes. As a result, further explorations are underway to determine the best alternative for a more informative and more precise representation.

Example 7. Three books.

TC: หนังสือ 3 เล่ม
(naŋsyˇy-book) (saˇam-three) (lêm-clas)

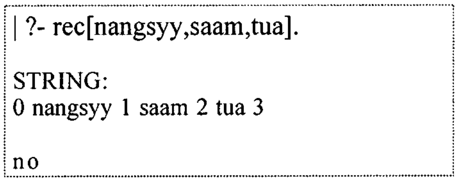
CT: หนังสือ 3 เล่ม



```
| ?- rec[naŋsyˇy,saˇam,lêm].

STRING:
0 naŋsyˇy 1 saˇam 2 lêm 3

CATEGORY:

phrase
SYNSEM synsem
    LOC loc
      CAT cat
         COMPS e_list
         HEAD noun
            AKO a12515
            CASE case
            MOD none
            PRD boolean
         MARKING marking
         PREV list_synsem
         SPR list_synsem
         SUBJ list_synsem
      CONT nom_obj
         INDEX [0] ref
            GEN neut
            NUM sg
            PER third
         RESTR ne_set_psoa
            ELT psoa
               NUCLEUS nangsyy
                  INSTANCE [0]
               QUANTS e_list
            ELTS e_set
```

```
| ?- rec[three,books].

STRING:
0 three 1 books 2

CATEGORY:

phrase
SYNSEM synsem
    LOC loc
      CAT cat
         COMPS e_list
         HEAD noun
            CASE case
            MOD none
            PRD boolean
         MARKING unmarked
         SPR e_list
         SUBJ e_list
      CONT nom_obj
         INDEX [0] ref
            GEN neut
            NUM pl
            PER third
         RESTR ne_set_psoa
            ELT psoa
               NUCLEUS books
                  INSTANCE [0]
               QUANTS e_list
            ELTS e_set
    NONLOC nonloc
       INHERITED nonloc1
          SLASH e_set
       TO_BIND nonloc1
          SLASH set_loc
```

FIGURE 11. Results of analyzing (a) "หนังสือ 3 เล่ม" and (b) "Three books."



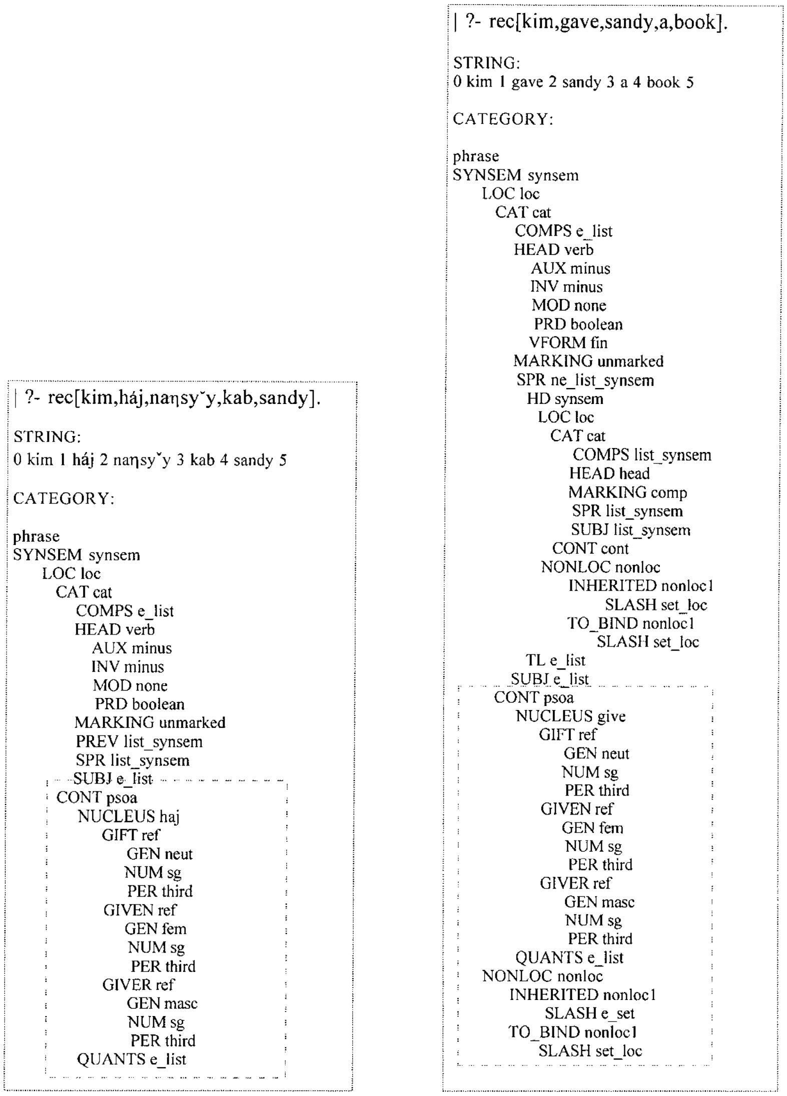
```
| ?- rec[nangsyy,saam,tua].

STRING:
0 nangsyy 1 saam 2 tua 3

no
```

FIGURE 12. Result of analyzing "หนังสือ 3 ตัว."

Example 8. Kim gave Sandy a book.

TC: คิม ให้ หนังสือ กับ แซนดี้
(kim-kim) (háj-give) (naηsyˇy-book) (kab-prep) (sandy-sandy)

CT: คิม ให้ หนังสือ กับ แซนดี้



```
| ?- rec[kim,háj,naηsyˇy,kab,sandy].

STRING:
0 kim 1 háj 2 naηsyˇy 3 kab 4 sandy 5

CATEGORY:

phrase
SYNSEM synsem
    LOC loc
      CAT cat
         COMPS e_list
         HEAD verb
            AUX minus
            INV minus
            MOD none
             PRD boolean
         MARKING unmarked
         PREV list_synsem
         SPR list_synsem
         SUBJ e_list
      CONT psoa
         NUCLEUS haj
            GIFT ref
               GEN neut
               NUM sg
               PER third
            GIVEN ref
               GEN fem
               NUM sg
                PER third
            GIVER ref
               GEN masc
               NUM sg
                PER third
         QUANTS e_list
```

```
| ?- rec[kim,gave,sandy,a,book].

STRING:
0 kim 1 gave 2 sandy 3 a 4 book 5

CATEGORY:

phrase
SYNSEM synsem
    LOC loc
      CAT cat
         COMPS e_list
         HEAD verb
            AUX minus
            INV minus
            MOD none
             PRD boolean
            VFORM fin
         MARKING unmarked
          SPR ne_list_synsem
            HD synsem
              LOC loc
                CAT cat
                   COMPS list_synsem
                   HEAD head
                   MARKING comp
                    SPR list_synsem
                    SUBJ list_synsem
                CONT cont
              NONLOC nonloc
                 INHERITED nonloc1
                      SLASH set_loc
                 TO_BIND nonloc1
                     SLASH set_loc
            TL e_list
          SUBJ e_list
      CONT psoa
         NUCLEUS give
            GIFT ref
               GEN neut
               NUM sg
               PER third
            GIVEN ref
               GEN fem
               NUM sg
                PER third
            GIVER ref
               GEN masc
               NUM sg
                PER third
         QUANTS e_list
    NONLOC nonloc
       INHERITED nonloc1
          SLASH e_set
       TO_BIND nonloc1
          SLASH set_loc
```

FIGURE 13. Result of analyzing (a) "คิมให้หนังสือกับแซนดี้" and (b) "Kim gave Sandy a book."

5. REPAIR AND ITERATE

There are two possible reasons why the generated TC may not retain the meaning of the SL. QDMT may select an inappropriate word for the input word (e.g., see Table 5, sentence 5). Selected words could be misordered by QDMT (see Table 5, sentences 2 and 4). By analyzing the semantic results in the TCE phase, we identify the incorrect parts; for example, if a different TC is generated for Example 8 as follows:

แซนดี้ ให้ หนังสือ กับ คิม
(sandy-sandy) (háj-give) (naηsyˇy-book) (kab-prep) (kim-kim)

then the result of analyzing (Figure 14) shows that the "giver" is "feminine" and the "given" is "masculine." RI compares the semantic information (in the dashed box) of the analyzed TC (Figure 14) with that of the analyzed SL (Figure 13(b)). The result of comparison identifies that the "giver" and the "given" of the generated TC are misordered. In this case, RI repairs the TC by switching the "giver" and the "given," resulting in the following TC:

คิม ให้ หนังสือ กับ แซนดี้

TCE analyzes this repaired TC and compares the result of analyzing with the parsed SL (Figure13(b)) again. This time there is no difference between the semantic information of the two parsings, Figures 13(a) and 13(b). So the repaired TC "คิม ให้ หนังสือ กับ แซนดี้" is acceptable as the translation of "Kim gave Sandy a book."

The Repair and Iterate phase is in its initial design stages at this point. Our prototype results with QDMT and TCE lead us to believe that we can continue to improve and refine these phases. We also believe, in the spirit of QDMT and TCE, that simple strategies may be all that are required in the RI phase. RI is a generally well-accepted strategy in many computational paradigms. For the GRMT philosophy, articulation of compositional strategies holds promise.

6. QDMT AND A COMMERCIAL MT SYSTEM VISIT A FEW SENTENCES

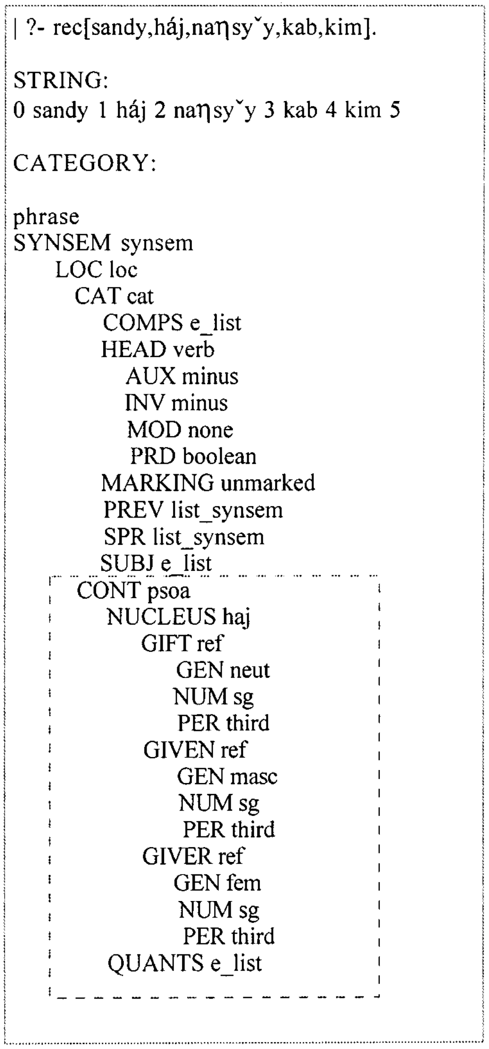
It is not our intention to compare QDMT to a commercial MT system. Rather, we illustrate how QDMT, with a simple application of constraints and principles, can obtain impressive results, obtaining TCs for subsequent processing. We have designed the application of constraints and the use of semantic principles to keep within the spirit of modern unification-based approaches to language analysis (e.g., HPSG, GPSG), using appropriate information when needed and subscribing to the general principle of compositionality of meaning.

In Table 7, each sentence is shown in five iterations: the original sentence in English, the translation in French of the original sentence as provided by a commercial MT system, the correct translation in French, the Thai translation candidate that was generated by QDMT, and the correct translation in Thai.

With the exception of the first sentence in Table 9 each generated French translation is incorrect in both word selection and grammar, and each generated TC is close to or the same as the correct translation.

7. CONCLUDING REMARKS

Generate and Repair Machine Translation is composed of three phases: quick and dirty machine translation, translation candidate evaluation, and repair and iterate. QDMT generates the translation candidate in a simple, straightforward manner, similar to the Direct approach,



```
| ?- rec[sandy,háj,naŋsyˇy,kab,kim].

STRING:
0 sandy 1 háj 2 naŋsyˇy 3 kab 4 kim 5

CATEGORY:

phrase
SYNSEM synsem
    LOC loc
      CAT cat
         COMPS e_list
         HEAD verb
            AUX minus
            INV minus
            MOD none
            PRD boolean
         MARKING unmarked
         PREV list_synsem
         SPR list_synsem
         SUBJ e_list
      CONT psoa
         NUCLEUS haj
            GIFT ref
               GEN neut
               NUM sg
               PER third
            GIVEN ref
               GEN masc
               NUM sg
               PER third
            GIVER ref
               GEN fem
               NUM sg
               PER third
         QUANTS e_list
```

FIGURE 14. Result of analyzing "แซนดี้ให้หนังสือกับคิม."

but more efficiently since QDMT accounts for differences between language pairs in terms of both syntax and semantics and this analysis ensures that the generated TC is exact or close to the correct translation.

The TCE analyses the translation candidate to determine if it conveys the meaning of the original sentence. If the TC does not, RI will repair it. These two stages, TCE and RI, ensure accuracy of the translation. They also ensure accuracy of translations from the TL to the SL. TCE and RI processes solve the problem of losing information during the transfer process of the Transfer approach, as was mentioned in Section 1.

GRMT treats the SL and TL separately, as is also the case of the Interlingual approach. GRMT is also aware of the differences between languages. Therefore, if languages can be grouped according to the various characteristics they have in common—for example,

TABLE 7. Commercial MT System Contrasted with QDMT.

1. The wheat was yellow.

   Le blé était jaune.
   Le blé était jaune.

   ข้าวสาลี นั้น สีเหลือง
   ข้าวสาลี นั้น สีเหลือง

2. When I was an ugly duckling he thought I never dreamed I could be so happy.

   Quand ètais un caneton laid, il pensait, je n'ai jamais rêvé que je pourrais être si heureux.
   Jamais je n' aurais rêvé , lorsque j' étais un rilain petit canard, que je pourrais être si heureux, pensa-t-il.

   เมื่อ ฉัน เป็น ลูกเป็ด ตัวหนึ่ง เขา คิด ฉัน ไม่เคย ฝัน ฉัน สามารถ เป็น ความสุข ขี้เหร่ มาก
   เมื่อ ฉัน เป็น ลูกเป็ด ขี้เหร่ ตัวหนึ่ง เขา คิด ฉัน ไม่เคย ฝัน ฉัน สามารถ มี ความสุข มาก

3. You can take a fish to school but you cannot make them think.

   Vous pouvez prendre un poisson pour scolariser, mais vous ne pouvez pas les faire penser.
   Vous pouvez emmener un poisson á l' école cependent il est impossible gue vous l'obligez de penser.

   คุณ สามารถ เอา ปลา ตัวหนึ่ง ไป ฝูง แต่ คุณ ไม่ สามารถ ทำให้ พวกเขา คิด
   คุณ สามารถ เอา ปลา ตัวหนึ่ง ไป โรงเรียน แต่ คุณ ไม่ สามารถ ทำให้ พวกเขา คิด

4. Five days before the trout are released.

   Cinq jours avant la truite sont publiis.
   Cinq jours avantque la truite son relâchée.

   ห้า วัน ก่อน ที่ ปลาเทราท์ นั้น จะ ถูก ปล่อย
   ห้า วัน ก่อน ที่ ปลาเทราท์ นั้น จะ ถูก ปล่อย

auxiliary verb, continuous tenses, passive voice—then the translation between groups can be performed more simply by GRMT.

QDMT has been implemented to generate TCs into Thai for English. Our initial experiments show that QDMT can generate the most appropriate TC for input sentences quickly and with relative accuracy. In many cases the translation is accurate without the need for subsequent processing. QDMT performs unsophisticated analyses efficiently. It is expected that the latter phases may be carried out in a similar manner.

Another aspect of concern in designing a machine translation system is the structure of the knowledge base—for example, the constraints, the AKO information in the dictionary. The structure of each knowledge base component should be direct, intuitive, and easy to extend for a large-scale MT system. In generating a reliable TC, QDMT requires simple information as illustrated throughout Section 3. This simplicity ensures that knowledge bases required in the QDMT phase are easy to manage in a large-scale MT effort. The dictionary used in earlier reported experiments of QDMT (Naruedomkul and Cercone 1997) was doubled in size to 378 English words and 570 Thai words. The "semantic relationship" and the "ordering rule" were also updated to serve the new dictionary. There has been no appreciable increase in processing time or storage structures when running QDMT on this larger dictionary. The information that is needed in the TCE and the RI phases has yet to be finalized; however, we are aware of the scalability problem in designing knowledge bases for these phases.

ACKNOWLEDGMENTS

The authors are members of the Institute for Robotics and Intelligent Systems (IRIS) and wish to acknowledge the support of the Networks of Centres of Excellence Program of the Government of Canada, the Natural Sciences and Engineering Research Council, and the participation of PRECARN Associates Inc.

REFERENCES

BROWN, P. F., A. D. P. STEPHEN, J. D. P. VINCENT, J. D. LAFFERTY, and R. L. MERCER. 1992. Analysis, statistical transfer, and synthesis in machine translation. *In* Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montreal, pp. 83–100.

CARPENTER, B., and G. PENN. 1994. ALE: The Attribute Logic Engine User's Guide. Philosophy Department, Carnegie Mellon University.

CERCONE, N. 1975. Representing natural language in extended semantic networks. Ph.D dissertation in Computing Science, The University of Alberta, Edmonton.

CICC. 1995a. Indonesian concept dictionary. Technical Report 6-CICC-MT65, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

CICC. 1995b. Indonesian generation rules. Technical Report 6-CICC-MT39, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

CICC. 1995c. Interlingual final edition. Technical Report 6-CICC-MT36, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

CICC. 1995d. Thai basic dictionary. Technical Report 6-CICC-MT55, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

CICC. 1995e. Thai concept classification. Technical Report 6-CICC-MT60, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

CICC. 1995f. Thai generation rules. Technical Report 6-CICC-MT50, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

GOODMAN, K., and S. NIRENBURG. 1991. The KBMT project: A Case Study in Knowledge-Based Machine Translation. Morgan Kaufmann Publishers, San Mateo, California.

HASS, M. R. 1964. Thai-English Student's Dictionary. Stanford University Press, Stanford, California.

JONES, D. 1996. Analogical Natural Language Processing. UCL Press Limited, London.

LANDSBERGEN, J. 1987. Isomorphic grammars and their use in the ROSETTA translation system. *In* Machine Translation Today: The State of the Art. Edinburgh University Press, Edinburgh.

MATHESON, C. 1996. HPSG grammars in ALE. "http://www.ltg.hcrc.ed.ac.uk/projects/ledtools/ale-hpsg/."

MITAMURA, T., E. H. NYBERG III, and J. G. CARBONELL. 1991. An efficient interlingua translation system for multi-lingual document production. *In* Proceedings of Machine Translation Summit III, Washington D.C., July.

MURAKI, K. 1989. PIVOT: Two-phase machine translation system. *In* Proceedings of the Second Machine Translation Summit, Tokyo, Omsha Ltd.

NARUEDOMKUL, K., and N. CERCONE. 1997. Steps toward accurate machine translation. *In* Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, Santa Fe, New Mexico, pp. 63–75.

PUNMETHA, N. 1984. Thai Grammar. Rungraungsarn (รุ่งเรืองสาสน์การพิมพ์) Ltd., Bangkok.

SAG, I. A., and T. WASOW. 1997. Syntactic Theory: A Formal Introduction. (Forthcoming).

SETHAPUTRA, S. (1977) New Model English-Thai Dictionary. Thai Wattana Panit Ltd., Bangkok.

TANTISAWETRAT, N., and SIRINAOVAKUL, B. 1991. An electronic dictionary for multilingual machine translation. *In* Proceedings of the Symposium on Natural Language Processing in Thailand, Chulalongkorn University, pp. 377–402.

UCHIDA, H. 1989. ATLAS-II: A machine translation system using conceptual structure as an interlingua. *In* Proceedings of the Second Machine Translation Summit, Tokyo, Omsha Ltd.