泰語機譯導向之控制性中文書寫規範研究

羅斌豪　史宗玲

摘要

由於愈來愈多的外籍勞工來到臺灣工作，其中又以泰國人為大宗，故我們開始重視與外籍勞工的有效溝通，且讓他們熟悉臺灣文化以適應新的環境變得日益重要。儘管外籍勞工能透過網路上的資訊來了解臺灣，但多數網頁都以中文或英文書寫，而並未以他們的母語書寫。若我們能將網路文本改用控制性中文來書寫，並使用線上 Google 軟體將其譯成泰語，泰國人則更能夠了解臺灣的文化。有鑒於此，本篇論文擬計劃運用機器翻譯理論、控制性語言理論及中文、泰文、英文這三種語言特性的差異作為理論框架，以探究：(1) 為泰語機器翻譯設計之控制性中文書寫規範及 (2) 分別為泰語與印歐語系語言所設計的控制性中文規範之異同。該研究使用專為印歐語系語言所設計的控制性中文規範（史宗玲，2011）為對比研究之準繩。此外，我們也邀請了十位泰國人來回答我們的問卷調查，以了解泰國人對於控制性文本及非控制性文本的泰語機器譯文之理解差異。

研究結果發現書寫泰語的控制性中文與書寫印歐語系的控制性中文有四大相似之處，包括：(1) 皆採用主語（主詞）加賓語（動詞）加謂語（受詞）的結構，(2) 皆使用目標語來取代文化相關的詞彙，(3) 皆使用文言文取代白話文，(4) 皆使用表不同時態的標記詞。另一方面，我們同時也發現書寫泰文的控制性中文與書寫印歐語系的控制性中文也有其相異之處，例如：(1) 在詞彙層面上，前者偏好使用雙音節字詞而後者則沒有

嚴格的規範，(2) 在句構層面上，前者偏好使用主動句而後者偏好被動句，(3) 在文法層面上，前者甚少使用不定冠詞、泰半使用量詞，而後者則偏好使用不定冠詞，卻甚少使用量詞。問卷回收後的結果發現，泰國人對於非控制性文本的泰語翻譯之理解程度平均為 42%，而對於控制性文本的泰語翻譯之理解程度則大幅提升到 86.2%。所有的受訪者皆認為控制性中文的書寫對泰語機器翻譯很有幫助。

綜而言之，研究的結果證實控制性中文之機器譯文品質可大幅改善。由此可知，我們可以針對不同的目標語來設計不同的控制性書寫規範，以改善其機器譯文的品質。最後，我們希望未來的學者與研究人員能夠針對不同的語組，如：日語與韓語，越南語與泰語等，繼續探討與設計出不同的控制性語言書寫規範，以全面提升網路語言即時溝通的成效。

關鍵詞：控制性中文的書寫規範、泰語機器翻譯、機器翻譯的品質

羅斌豪，國立高雄第一科技大學應用英語系口筆譯碩士生。

史宗玲，國立高雄第一科技大學應用英語系教授。

本文部分初稿內容曾發表於「2013 臺灣翻譯研討會 —— 科技與翻譯發展」，作者在此向該場研討會主持人董大暉教授、評論人史宗玲教授及所有與會人士，以及本文的兩位匿名審查者所給予的修正意見致謝。

Normative Design of Controlled Language Writing for Thai MT Application

Pin-Hao Lo　Chung-Ling Shih

Abstract

As an increasing number of foreign workers, particularly the Thai people, come to work in Taiwan, we have paid more attention to effective communication with the Asian foreigners. It is an urgent concern to make these people adapt to a new environment by familiarizing them with Taiwan's culture. Although they can learn about Taiwan by reading information on the Internet, the web texts are mainly written in Chinese or in English, not in their native languages. Thus, we consider controlling the source Chinese text and using Google Translate to translate it into Thai language so that Thai people in Taiwan can learn about Taiwan's culture and relevant others through Thai machine translation. Drawing on the theories of machine translation (MT), controlled language writing and contrastive analysis of the linguistic features of Chinese, Thai and English as the theoretical framework, this paper aims to investigate what norms should be followed in controlled Chinese (CC) writing for effective Thai MT application and how the norms are different from those used for Indo-European MT application. Three cultural texts on Ghost Festival, Beehive Fireworks and Qixi, and two texts on tourist attractions of Sun Moon Lake and Yehliu Queen's Head are adapted in CC for Thai MT application. The CC norms, particularly designed for the machine-created translations in Indo-European languages proposed by Shih (2011), are used as a yardstick to identify some similarities and differences between CC norms for Thai MT and for MTs in western languages. In addition, a questionnaire was administered on ten Thai native speakers to know the Thai audiences' reception of the Thai MT outputs of both texts in uncontrolled Chinese and in CC respectively.

The findings showed that the similarities between the CC norms for Thai MT and for MTs in western Indo-European languages included: 1) the use of SVO structure, 2) the use of cultural terms in target languages, and 3) the use of vernacular Chinese, not classical Chinese and 4) the use of time markers to indicate different verb tenses. The similar use of the SVO structure results from the linguistic features commonly shared by Thai language and English. In contrast, there are some different norms, including: 1) the use of monosyllabic/multisyllabic words in the lexical area, 2) the use of active and passive voice in the syntactic area, and 3) the use of indefinite article and measure words in the grammatical area. With regard to the results of the questionnaire, we found that Thai respondents' understanding of the Thai MT output of the uncontrolled Chinese text was 42% on average while that of CC texts rose up to 86.2% on average. All the Thai respondents agreed on the effectiveness of CC use for Thai MT application.

In a nutshell, this research suggests that CC norms can be designed for different target languages through MT application and therefore the understandability of the MT output can be highly improved. With a set of CC norms tailored for Thai MT application, language barriers could be overcome and Thai people can access more information about Taiwan. In light of the benefits of CC writing, we recommend that the similar research model be applied to the design of different controlled languages for the effective MT application in different source and target language pairs, including Vietnamese and Thai, Japanese and Korean, and relevant others in future studies.

Keywords: Controlled Chinese norms, Thai MT application, improved quality of MT outputs

Pin-Hao Lo, MA student, T&I MA Program, National Kaohsiung First University of Science and Technology

Chung-Ling Shih, Professor, the English Department and T&I MA Program, National Kaohsiung First University of Science and Technology

1. Introduction

An increasing number of foreign workers, especially the Thai people, come to Taiwan, get married with Taiwanese and settle down in Taiwan. It is thus important to introduce Taiwanese culture to this group of people. However, due to the language barrier, they might not be able to easily access the online information written in Chinese. Even though people can get the gist of information written in Chinese with the help of online Google Translate, the output of the Thai MT is often far from being understood well. To improve readability and comprehension of the Thai MT, this research proposes writing Chinese controlled texts for the Thai machine translation created by Google Translate on the web. It is assumed that the use of controlled Chinese text can improve the Thai MT performance and boost Thai audiences' understanding of online cultural information.

In Taiwan, Thai people often encounter difficulty understanding the public signs and the tourist information written in Chinese, English or Japanese in many tourist attractions. For this reason, if we edit the source texts of Taiwan's tourist attractions in controlled Chinese and improve the quality of their Thai MT, the Thai people could easily understand the information about the tourist attractions. In this respect, to help the Thai people understand Taiwan's local culture and tourist attractions through the Thai MT on the web, this paper will investigate how to re-author the Chinese texts in controlled language for effective Thai MT application. Simply put, this paper aims to identify some norms for customizing the controlled Chinese texts to create understandable the Thai MT output.

When it comes to the subject of controlled language, many scholars have conducted research on it. For example, in an online article entitled "Controlled Language", Richard Nordquist discussed some text types suitable for MT application and cited John Kirkman's (1992) statement that controlled English is mostly used to author the business and technical texts such as installation instructions, maintenance and repair procedures and relevant others. In an article entitled “From Plain English to Controlled English”, Diego Mollá and Rolf Schwitter proposed two steps and some rules for writers to follow in writing technical texts in controlled English (CE) for MT application. The two steps include: 1) analysis of the text and its translation based on a set of logical forms; 2) determining the components in controlled language texts. In addition to the two main steps, they also summarized four rules for CE writing, including: 1) “prepositional phrases in adjunct position always modify the verb; 2) relative sentences modify the immediately preceding noun; 3) only of-constructions are allowed as post-nominal modifiers and 4) abbreviations are resolved to their full form.” (Schwitter 1998, p.4)

In another article entitled “SMART Controlled English – Paper and Demonstration”, John M. Smart suggested three steps for people to learn the writing in controlled English, including: 1) sorting out good terms and store them in the database; 2) checking the grammar rules to form grammatically correct sentences, and 3) receiving a training in becoming a good controlled language writer. (John M. Smart, 2006) In a book entitled *Real-time Communication Through Machine-enabled Translation: Taiwan’s Oracle Poetry*, Shih, Chung-ling (2011) proposed three controlled Chinese (CC) strategies that are customized for machine-created translation in Indo-European languages. These strategies include simplification, normalization and explicitation and she used Taiwan’s oracle poetry as a case study.

The papers and book introduced above show a focus on the study of writing CE and CC for the improved MT performance. Unlike previous studies, this paper aims to identify some rules for CC writing specifically designed for the Thai MT, and use them to measure against the rules proposed by previous scholars. Thus, we can identify the similarities and differences between the CC writing norms for Thai MT and those for Indo-

European languages MT.

To achieve the purpose of this research, three research questions (RQs) are raised for investigation, including:

1) What norms do Chinese controlled texts have to follow to improve the semantic and grammatical accuracy of the Thai MT?
2) How are the Chinese-controlled norms for the Thai MT similar to or/and different from the norms for the western language MT?
3) What is the Thai audiences' reception of the Thai MT of CC texts?

The findings of RQ1 can be used as a yardstick to measure against previous norms proposed by some scholars. The answer to RQ2 helps us understand the similarities and differences between CC writing for Thai and for western languages. By doing so, future researchers or CC text writers can take different approaches and strategies based on the finding while writing CC texts for the translation created by online Google Translate in different target languages. The answer to RQ3 aims to show that the understandability of the Thai MT output can be enhanced after we learned the Thai audiences' reception and revised the norms for CC writing.

2. Theoretical Review

The present research aims to identify a few norms for CC writing customized for the Thai MT application by measuring them against those proposed by previous scholars for the MTs in Indo-European languages. Consequently, controlled language theory and MT/ Google Translate will be used as the analytical framework to support the arguments in this research. Moreover, to analyze the different CC norms designed for different target languages, a comparison among Chinese, Thai language and English will be given. The following paragraph will introduce some basic concepts of controlled language theory, MT system/Google Translate and a comparison of linguistic features among Chinese, Thai language and English.

2.1 Controlled language theory

Writing a text in the controlled language means an execution of "some limitations in the scope of vocabulary and sentence patterns in the source text" with an aim to improve the MT output (Shih, p. 202). The limitations suggest the use only single-meaning, common, unambiguous words and the short sentences of simple construction. Writing CL texts can reduce the difficulty in understanding a sentence because CL texts use fewer, if not none, complicated sentences. Writing controlled texts can also reduce the ambiguity of the MT output by using single-meaning and common words. Controlled English writing has been applied to MT-mediated multilingual translations of software, user manuals and manufacturing documentation. One of the examples is Boeing Company. They once funded a Simple English project that asked technicians to author technical texts in CE. To date, CE has been applied to the localization industry. Rolf Schwitter, Anna Ljungberg and David Hood (2003) proposed that the languages could be controlled from two aspects, grammar and lexicon. Some basic rules for writing CE texts include the use of active voice, the use of simple verb tense, the consistency in the choice of words, the use of single-meaning words and the use of shorter sentences. These are all useful and important norms to be followed.

CE has been long discussed for years. Recently, an increasing number of scholars have turned to focus on Controlled Chinese (CC) writing and one of them is Shih, Chung-ling. In her book *Real-time Communication through Machine-enabled Translation: Taiwan's Oracle Poetry*, she proposed several CC norms suitable for the MT of Indo-European languages, such as French, German and Spanish, just to name a few. She proposed three strategies for writing Taiwan's oracle Poetry in CC, including simplification, normalization and explicitation. Moreover, she provided some specific rules to support them. The rules included the use of vernacular Chinese, not classical one in the lexical area, the use of the SVO structure and the passive voice in the syntactic area, the use of "jie" [皆] plus nouns to show the plural forms in English in the grammatical aspect and the use of English names for cultural terms in the pragmatic aspect. The purpose of applying these rules is to boost the communication effectiveness of the MT outputs.

In writing CL texts, stricter rules need to be applied. However, even though writing CL texts helps boost the readability and improve the semantic accuracy of the MT, there is an underlying drawback. It is that no fixed set of rules can be applied to all languages in the world. CL rules must be designed for different target languages. In this paper, a set of CL rules for the Thai MT will be identified and its difference from those tailored for Indo-European languages MTs will be discussed.

2.2 MT/ Google Translate

MT is the abbreviation of Machine Translation, also known as Computer Translation. It was first used in the 1960s. MT can be briefly divided into two types in terms of the number of languages involved, bilingual and multilingual systems. The former deals with the translation between two languages, whereas the latter copes with a translation from one source language into several target languages. The MT tool is mainly divided into three types: rule-based, statistics-based and example-based ones. Rule-based MT tries to identify the grammar rules of ST and translates it into TT based on the grammatical rules. The statistics-based one basically operates using bilingual corpora. Example-based one is to "translate a source sentence by imitating the translation of a similar sentence already in the database" (Sago

& Nagao, 1990:1).

Unlike the above MT tools/systems, Google Translate is corpus-driven. The working process is to retrieve the highly matched segments from the corpora and string them up to give the translation. So far, Google Translate is more suitable for Ch-to-En translation than En-to-Ch. Google Translate is accessible to the public for free. Thus, all MT tests in this paper use Google Translate.

Though Google Translate is getting more popular in the technological era, it is still facing some criticism, such as low quality, literal translation and wrong wording. Compared with translation done by humans, the output of MT is far from being comprehensible. MT cannot process well the long sentences and the words with ambiguous meanings or multiple meanings. To overcome this defect, it is necessary to write texts in the controlled language to help improve the quality of their MT outputs.

2.3 The differences in linguistic features among Chinese, Thai and English

As this research will measure CC norms for the Thai MT against the norms proposed by previous scholars for Indo-European languages, it is necessary to understand the differences among Chinese, Thai language and English. These three languages belong to different language families. Chinese belongs to Sino-Tibetan language family; Thai, Tai-Kadai language family, and English, Indo-European language family. Due to the different language families they belong to, they embody similar and different linguistic features. After consulting some information (Savetamalya, 1996; Luo & Sun, 2003; Mixdorff, Charnvivit & Luksaneeyanawin, 2006; Zhang, 2006; Duanmu, 2007; Yaowapat & Prasithrathsint, 2009; Yang, 2011; Jenks & Huang, 2012), we summarized a number of similarities and differences among Chinese, Thai and English. Table 1 shows the different linguistic attributes of Chinese, Thai and English.

Table 1. ***Linguistic attributes of Chinese, Thai and English***

	Chinese	Thai	English
Linguistic Systems	Sino-Tibetan Language	Tai-Kadai Language	Indo-European Language
Passive Voice	Not frequently used Often used to describe the negative situation	Not frequently used Often used to describe the negative situation	Frequently used
Structure	SVO	SVO	SVO
Classifiers/ Measure Words	Yes	Yes	Only certain nouns
Serial Verb Construction	Yes	Yes	No
Tense Marker	Yes	Yes	No
Singular/ Plural Form	No	No	Yes
Two-character Words/ Disyllabic Words	More frequently	More frequently	Less frequently
The Position of Classifiers	Numbers + Classifiers + Nouns	Nouns + Numbers + Classifiers	Numbers + Classifiers + Nouns
The Position of Modifiers	Pre-modification + Nouns	Nouns + Post-modification	Nouns + Post-modification

As shown in Table 1, Chinese and Thai languages are both Asian languages, thereby sharing more linguistic features than Chinese and English do. In the syntactic area, both of Chinese and Thai do not frequently use passive voice, and the passive voice is often used to describe the negative situations such as punishment, car accident or death. Additionally, both of them tend to use the subject + verb + object (SVO) structure; both of them have the serial verb construction. That is, several verbs are allowed to be used in a sentence. For instance, we may say: 我打開門走進來拿書 (lit: I open the door coming inside for books) in Chinese and ผมเปิดประตูเข้ามาเอาหนังสือ in Thai (lit: I open the door coming inside for books). In the grammatical area, both of them tend to use classifiers and measure words. Furthermore, both languages use tense markers to show various tenses in a sentence, such as 了 and 過 in Chinese or แล้ว and เคย in Thai. In addition, the plural form of nouns is the same as the singular form in both languages. Finally, in terms of wording, people tend to use two-character words in Chinese and disyllabic words in Thai. Several studies have shown that Chinese native speakers tend to use two-character words, especially in writing. Moreover, in a statistical analysis of the frequencies of using Thai words, two syllables words predominate, accounting for 41.9%, compared with monosyllabic words, 21.4% and three syllable words, 21%.

Even though Chinese and Thai share some similar linguistic features, both still have some differences. Two examples may be given, including the positions of classifiers and modifiers. In Chinese, the use of classifiers is often structured as “number + classifier + noun”, such as 「一本書」(a book), while in Thai, “noun + number + classifier” such as หนังสือหนึ่งเล่ม (a book). Moreover, Chinese uses “modifier + noun”, such as 我看過的書 (lit: I read de book/ the book that I read), while Thai uses the structure, “noun + modifier, ex: หนังสือที่ผมเคยอ่าน (lit: the book that I read).

In contrast, English is quite different from Chinese and Thai in the categories analyzed above. English tends to use passive voice to express an idea, either positive or negative. In terms of the use of classifiers, only few nouns are paired with specific classifiers, such as a pair of glasses or a piece of paper. Otherwise, the number is only followed by nouns and carries no classifiers. Additionally, there is no serial verb construction in English. When several verbs form a sentence, the rest of the verbs need to convert their forms into gerunds. English does not rely on tense markers to indicate the tense. The verb needs to be conjugated in accordance with different tenses. Furthermore, in English, nouns are often added “s” or “es” to show the plural form. Lastly, Chinese and Thai use many compound words, while English only uses few.

Drawing on some basic concepts of controlled language theory, MT/Google Translate and the linguistic features of Chinese, Thai language and English as the theoretical framework, this paper will investigate a set of CC rules for the Thai MT application. Most importantly, a comparison among those three languages aforementioned clearly shows their different linguistic features and helps to identify different CC norms applied to various target languages.

3. Methodology

3.1 Data

To fulfill the objective of making Thai people easily access the information of Taiwan’s festivals and famous tourist attractions, this paper uses the CC texts on the relevant topics for effective Thai MT application. The two texts on famous tourist spots in Taiwan introduce Yehliu and Sun Moon Lake. The three web texts about Taiwan’s festivals introduce the Ghost

Festival, Yensui's Beehive Fireworks and the Qixi. The internal structure of collected data is tabulated as Table 2 shows:

Table 2: ***The collected data for adaptation in CC***

No	Text Titles	Words Counts	URL/ Website address
1.	The Ghost Festival [中元節]	562	http://taiwanpedia.culture.tw/web/content?ID=11714
2.	Yanshui's Beehive Fireworks [鹽水蜂炮]	512	http://taiwanpedia.culture.tw/web/content?ID=4470
3.	Qixi (七夕)	507	http://taiwanpedia.culture.tw/web/content?ID=11707
4.	Sun Moon Lake [日月潭]	168	http://www.sunmoonlake.gov.tw/AboutSunmoonLake/about01.htm
5.	Yehliu Queen's Head [野柳女王頭]	219	http://www.ylgeopark.org.tw/content/landscape/Sight.aspx

The source texts about festivals can be found in *Encyclopedia of Taiwan*(臺灣大百科全書). As for texts about tourist attractions, the text of Sun Moon Lake is extracted from the website of Sun Moon Lake National Scenic Area(日月潭觀光旅遊網) and the text of Queen's Head(女王頭) is retrieved from the website of Yehliu Geopark(野柳地質公園全球資訊網). The hyperlinks are offered in Table 2 for the readers' reference.

3.2 Methods

3.2.1 Chinese controlled writing and MT tests

After collecting online information from websites, we re-authored the texts in controlled Chinese. Next, the CC texts were sent to Google Translate for the Thai MT test. To improve the quality of the Thai MT, several adjustments were needed. In doing so, certain norms for writing CC texts to improve the accuracy of Thai MT were detected. The norms retrieved from the steps aforementioned will be further measured against the norms for Indo-European languages proposed by previous scholars to identify their similarities and differences.

3.2.2 A questionnaire

This research aims to prove that the quality of the Thai MT output can be greatly improved as long as the Chinese ST is carefully controlled. Moreover, the paper will detect some CC norms for Thai MT application. To achieve the first objective, a questionnaire is designed and done by Thai native speakers. The interviewees are expected to be ten Thai native speakers in total and each two are invited to read the Thai MT of the same CC text. The questionnaire consists of two parts. The first part is an evaluation on the Thai MT outputs of natural Chinese texts, and the second part, an evaluation on the Thai MT outputs of CC texts. Each part raises three open questions. They are: 1) How much can you understand the text in terms of the clarity of meaning and grammatical accuracy? (Give us a percentage number). 2) What is the main factor that prevents you from understanding the text? (nouns, verbs, cultural terms or sentence structure) Please choose one. And 3) based on your answer on Q2, give us one supportive example. The answer to the first question is expected to reveal how much native Thai speakers can understand the MT output and how much can be improved after the adaptation of the text in CC by comparing the percentage of comprehension obtained from each section. The second question helps us understand what plays a crucial part in causing Thai people's difficulty of understanding the Thai MT. By examining the answer, we can infer what linguistic components, such as nouns, verbs, and cultural terms, and what sentence patterns have confused the Thai readers most. The third question provides a specific example to illustrate what really hinders their comprehension.

3.3 Analysis criteria: CC Norms/ rules

Norms retrieved from this research will be used to measure against norms proposed by Shih (2011) in her book *In Real-time Communication through Machine-enabled Translation: Taiwan's Oracle Poetry*. As mentioned above, the rules that are designed for Indo-European languages MT serve as research criteria to help us identify the similarities and differences between CC norms for the Thai MT and for the Indo-European languages MT. Some major rules are chosen from Shih's book as benchmarks. Table 3 shows the chosen CC norms as research criteria in this paper.

Table 3: ***The chosen CC norms for Indo-European languages MT***

	CC Rules
Lexical Area	The use of vernacular Chinese, not classical Chinese
Syntactic Area	1. The use of the SVO structure 2. The use of passive voice
Grammatical Area	1. Adding the indefinite article and measuring words 2. The use of "dangshi"[當時] before a verb to show the past verb tense 3. The use of "gai"[該]to suggest 'the'
Pragmatic Area	1. The use of the English names of cultural terms 2. No use of superstitious, overlapping and inappropriate information

The above rules will be used to measure if they are applicable to the controlled text for effective Thai MT. And the rules would be examined if some of them should be deleted and if some new rules should be supplemented.

4. Findings and Discussions

In this section, the paper attempts to provide answers to previous research questions raised in Section One and some inferences will be given after the analysis of the questionnaires done by the respondents, ten Thai native speakers.

4.1 The identification of CC norms for Thai MT application

In response to RQ1 about what norms Chinese controlled texts have to follow to improve the semantic and grammatical accuracy of the Thai MT, some CC norms for the Thai MT in lexical, syntactic, grammatical and pragmatic areas were identified. After we tested the CC texts following Shih's norms on Google Translate, and obtained Thai people's responses to the Thai MT outputs, the norms are tabulated as Table 4 shows.

Table 4: ***The CC norms for the Thai MT***

	CC norms
Lexical Area	1. The use of vernacular Chinese, not classical Chinese 2. The use of disyllabic or multisyllabic words for Thai disyllabic words
Syntactic Area	1. The use of the SVO structure 2. The use of active voice 3. The use of transitional words or connecting words
Grammatical Area	The use of post-noun modifiers in accordance with Thai sentence structure
Pragmatic Area	1. The use of Thai names for cultural terms, the name of a person, a place or a proper noun 2. The retention of superstitious part in culture-related texts 3. Extra information and explanation for unfamiliar topics, such as a festival, the origin of a story, a brief history of a god and goddess

The following part gives some examples to show the differences in the MT outputs before and after the use of the CC text.

Ex1: Use of vernacular Chinese, not classical Chinese

ST: 後來玉帝感其至誠，乃特准他們在每年七夕夜相會。

CC:จักรพรรดิหยก 感覺深受感動。จักรพรรดิหยก 允許他們看到對方，在 7 月 7 日，每一年。

MT1:ต่อมาจักรพรรดิแห่งความรู้สึกของความจริงใจเป็นผู้มีอำนาจในคืนทานาบาตะประจำปีพวกเขาได้พบ

(lit: Late Emperor of the feeling of the seriousness is an authority at the night every year they met.)

MT2:จักรพรรดิหยกรู้สึกสะเทือนใจ จักรพรรดิหยกช่วยให้พวกเขาได้เจอกันอีกในวันที่ 7 กรกฎาคมของทุกปี

(lit: The Jade Emperor felt touched. The Jade Emperor helped them to meet again on July 7 every year.)

The norm applied here is the use of vernacular Chinese, not classical Chinese. So classical Chinese such as 感其至誠 and 特准 are replaced with vernacular Chinese 深受感動 and 允許 . The MT outputs of classical Chinese words ความรู้สึกของความจริงใจ (the feeling of the seriousness) and ผู้มีอำนาจ (an authority) are thus replaced with รู้สึกสะเทือนใจ (felt touched) and ช่วยให้ (helped) in the MT output of the CC text. By contrasting the MT1 and the MT2, we see that the MT1 is semantically incorrect but the MT2 is correct because its source text is written in CC.

Ex2: The use of disyllabic or multisyllabic words for Thai disyllabic words.

ST: 拜七娘媽的物品，大多是蔴油雞、油飯、香粉、香水、鮮花、針線、鏡子、金紙。

CC: 父母將會準備許多東西，包括雞的肉、糯米飯、香的粉、香水、花朵、針和線，且鏡子。

MT1:ขอบคุณรายการ Qiniangma ไก่งาส่วนใหญ่น้ำมัน, ข้าว, แป้ง, น้ำหอมดอกไม้เย็บปักถักร้อย, กระจก, กระดาษทอง

(lit: Thanks you things Qiniangma sesame chicken oil, rice, powder, perfume, flowers, embroidery, mirror, paper gold.)

MT2:พ่อแม่ผู้ปกครองจะเตรียมหลายสิ่งหลายอย่างรวมทั้งเนื้อไก่, ข้าวเหนียว, ผงหอม, น้ำหอม, ดอกเข็มและด้ายและกระจก

(lit: Parents will prepare many things include chicken, sticky rice, fragrant powder, perfume, needles and thread and mirrors)

The example shows that even though 油飯 and 香粉 in the source text are two-character words, sometimes multisyllabic words need to be used to produce disyllabic words in the Thai MT. Consequently, in the CC text, 糯米飯 (ข้าวเหนียว) and 香的粉 (ผงหอม) are used to replace 油飯 (ข้าว) and 香粉 (แป้ง) so that disyllabic words will be produced in the Thai MT output.

Ex3: Use of the SVO structure

ST: 普度時應備髮梳、鏡子、春仔花等。

CC: 一般民眾 (S) 也會放置 (V) 梳子、鏡子和花，用紙做成的 (O)。

MT1:เพอร์ควรจะเตรียมความพร้อมหวี, กระจก, ฤดูใบไม้ผลิดอกไม้อเบอร์ดีน

(lit: Perry will prepare readiness comb, mirror, spring flowers, Aberdeen.)

MT2:ประชาชนทั่วไปจะวางกระจก, หวีและดอกไม้ที่ทำจากกระดาษ

(lit: The public will place mirror, comb and flowers made of paper.)

The ST shows that Chinese sometimes follows the topic-comment sentence structure. In other words, 普度時 is the topic and 應備髮梳、鏡子、春仔花等 is the comment. However, the topic-comment sentence structure is hard for Google Translate to produce a correct translation. Consequently, the SVO structure is used in the CC text to help the MT system create the Thai translation that has clear meaning for the Thai audience to easily understand.

Ex4: Use of transitional words

ST: 公普是地方寺廟舉行法會，私普是指各行各業自行協調一天聚集普度。

CC: 在臺灣，我們會舉辦 เทศกาลพ้อต่อ 在廟裡。此外，許多機構及公司也可以安排 เทศกาลพ้อต่อ

MT1:ประชาชนทั่วไปจะจัดขึ้นที่วัดทั่วไปในพื้นที่ส่วนตัวของตัวเองบูชาหมายถึงทุกเดินชีวิตรวมเพอร์ดูการประสานงานหนึ่งวัน

(lit: The general public will hold in temples in private area of their own. Worship means all walks of life including Purdue to coordinate a day.)

MT2:ในไต้หวันเราจะมีการจัดเทศกาลพ้อต่อในวัด นอกจากนี้หลายสถาบันและ บริษัท ยังสามารถจัดเทศกาลพ้อต่อ

(lit: In Taiwan, we will organize Pudu in temples. Besides, many institutes and companies can also organize Pudu.)

The example shows that the output of the uncontrolled ST is incomprehensible. Moreover, by adding a transitional word [此外] in the CC text, logical relations between two clauses or sentences in the Thai MT are strengthened and thereby helps reduce the decoding burden of readers.

Ex5: Use of post-modification

ST: 環潭公路

CC: 公路，圍繞 ทะเลสาบสุริยันจันทรา

MT1:ถนน Central Lake

(lit: road Central Lake)

MT2:ทางหลวงรอบทะเลสาบสุริยันจันทรา

(lit: road around the Sun Moon Lake)

The example shows that Chinese follows the pattern of modifier + noun structure while Thai follows that of noun + modifier. To increase the readability and accuracy of the Thai MT output, we should conform to the Thai linguistic convention while writing CC texts for Thai MT.

Ex6: Use of cultural references in the target language
ST: 普渡、七夕、牛郎星、織女星
CC: เทศกาลพ้อต่อ ราตรีแห่งเลขเจ็ด ทะเลสาบสุริยันจันทรา ดาวอัลแทร์ ดาวเวกา
MT1: เพอร์ดู, ทานาบาตะ, ทะเลสาบ Sun Moon, Altair, Vega
(lit: Purdue, Tanabata, Lake, Sun Moon, Altair, Vega.)
MT2: เทศกาลพ้อต่อ ราตรีแห่งเลขเจ็ด ทะเลสาบสุริยันจันทรา ดาวอัลแทร์ ดาวเวกา
(lit: Pudu Festival, Qixi Festival, Sun Moon Lake, Altair, Vega.)

The example clearly shows that writing culture-related terms in the source language does not help Google Translate produce an understandable translation. Instead, writing CC texts in the target language greatly enhances the semantic accuracy of the Thai MT output.

Ex7: Addition of culture-specific explanation
ST: 居民到武廟祈求，迎請關聖帝君於元宵節出巡遶境。
CC: 此後，一般民眾會來祭拜 กวนกง，且請求 กวนกง 前去各地區視察每年。กวนกง 表現了正直與忠誠的精神。即使他死後，中國人依舊祭拜 กวนกง。
MT1:ชาวโมวัดสวดมนต์ทักทายควนขบวนแห่โคมไฟ Yu รอบชายแดน
(lit: The Mo Temple prayer greet Kuan Yu lantern parade around the border.)
MT2:ตั้งแต่นั้นมาประชาชนทั่วไปจะมานมัสการกวนกงและขอกวนกงไปตรวจสอบแต่ละภูมิภาคเป็นประจำทุกปี กวนกงแสดงให้เห็นถึงจิตวิญญาณของความซื่อสัตย์และความจงรักภักดีแม้หลังจากการตายของเขาคนจีนยังคงบูชา กวนกง
(lit: Since then, the public will come to worship Guan Gong and asked Guan Gong to visit each region annually. Guan Gong represents the spirit of honesty and loyalty. Even after his death, the Chinese still worship Guan Gong.)

The example shows that the names of culture-related persons, Gods, events or objects are unknown to foreigners, so extra explanations could be supplemented to increase the readability and understandability of their machine-created translations. In addition to the clear understanding of the contexts, the Thai audiences can gain more cultural knowledge.

4.2 Similarities and differences between CC norms for Thai MT and for Western languages MTs

In reply to RQ2 about the similarities and differences between the norms designed for the Thai MT and for western languages MTs, the findings show that the similarities include the use of vernacular Chinese, the use of the SVO structure and the use of the target languages for cultural terms. The use of vernacular Chinese, not classical Chinese, helps the MT system create the more accurate translation. Moreover, writing CC texts with the SVO structure reduces the ambiguity of the MT output. The use of the target language for cultural terms guarantees the accuracy of the MT output, reducing the risk of incomprehensibility in the MT output. Due to the shared common linguistic feature such as the use of the SVO structure in both Thai language and English, writing CC texts for both languages uses the SVO structure. However, the strategies for using the vernacular Chinese and the target languages for cultural terms have no much to do with linguistic features.

Even though the norms for the Thai MT and that for western languages MT share some similarities, there are also some differences between them, including: 1) the use of monosyllabic/multisyllabic words in the lexical area; 2) the use of active and passive voice in the syntactic area, and 3) the use of indefinite article and measure words and the words used for plural forms and past verb tense in the grammatical area. In the lexical area, as presented in the section of theoretical review, Thai uses many compounds words, while

English tends to use fewer compound words. For example, the CC text for the MTs in western languages uses 雞 (chicken)、粉 (powder) while the CC text for the Thai MT needs to use 雞的肉 (chicken meat)、香的粉 (fragrant powder) to produce Thai disyllabic words. Writing CC texts for the Thai MT thus needs to put compound words into consideration in order to produce disyllabic Thai words, making the Thai MT more reader-friendly.

In the syntactic area, Thai tends to use more active voice, while English prefers passive voice. For example, in writing CC texts for the Thai MT, we write 我們會邀請道士前來主持這儀式，以祭拜鬼魂 (lit: We will invite priests to preside over this ceremony to worship ghosts), while the CC texts for western languages MT goes as 僧侶和道士是被邀請來主持儀式，以祭拜鬼魂。(lit: Monks and priests were invited to preside over the ceremony to worship ghosts). Consequently the strategies for writing CC texts for Thai MT and for western languages MTs are different in some respects. In the grammatical area, since the Thai language is not a language which needs to show plural forms of nouns and the past verb tense as English is required, the grammatical norms of writing CC texts for Thai MT and for western languages MTs are thus varied. For example, we write「當時」(lit: at that time) in the CC texts for western languages MT to show the past verb tense, such as 該男人當時是一凡人，但那個女孩當時不是。Its English MT is *The men at that time was a mortal but the girl at the time was not*. In writing CC texts for the Thai MT,「當時」(lit: at that time) could be dropped and the sentence 男人那人是普通人，但那個女孩不是。Its Thai MT is ผู้ชายคนนั้นเป็นคนธรรมดา แต่ผู้หญิงคนนั้นไม่ได้เป็น. The example shows that Thai language needs not to present the past verb tense.

4.3 Thai audiences' reception

In response to RQ3 about the Thai audiences' reception of the Thai MT output of CC texts, the statistical results obtained from the questionnaires done by ten Thai native speakers showed that Thai respondents' understanding of uncontrolled Chinese texts was 42% on average while that of controlled Chinese texts rose up to 86.2% on average. This means that the Thai MT of the uncontrolled source text is far from being comprehensible, but that of the controlled text greatly improves. The percentage of how much the Thai audiences understand on average is given as figure 1 shows.

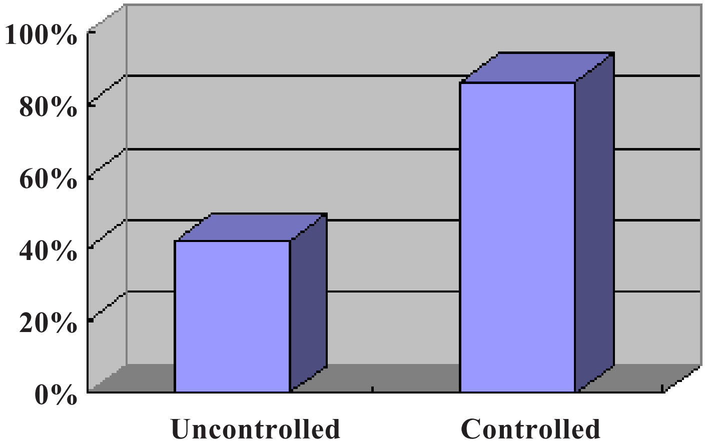

Figure 1: ***Thai audiences' different reception of the Thai MT outputs of uncontrolled and controlled Chinese texts***

Specifically put, in reply to the MT of the text on Yehliu, the respondent Jia expressed that she understood 40% of the uncontrolled text and 65% of the CC text. Ekk understood 70% of the uncontrolled text and 90% of the CC text. Regarding the MT of the text on Sun Moon Lake, the percentage of the understanding of the uncontrolled text for Nack was 40% and the CC text, 70%. Nad understood 50% of the uncontrolled text and 80% of the CC text. In the MT of the text on Ghost Festival, Tiger understood 70% of the uncontrolled text and 100% of the CC text. Korn understood 80% of the uncontrolled text, and 99% of the CC text. In the MT of the text on Yanshui

Beehive Fireworks, Pongarpa understood 10% of the uncontrolled text, and 100% of the CC text. Maem understood 0% of the uncontrolled text, and 98% of the CC text. In the MT of the text on Qixi, Pai and Gai both understood 30% of the uncontrolled text, and 80% of the CC text.

In short, the statistical survey of the questionnaire indicated that the quality of the Thai MT was greatly enhanced after the ST was carefully controlled using the norms particularly designed for the Thai language.

5. Conclusions

To sum up, the paper has identified a set of norms tailored for Thai MT application, including the use of vernacular Chinese and disyllabic or multisyllabic words in the lexical area, the use of the SVO structure, active voice and transitional words or connecting words in the syntactic area, and the use of post-modification to match the Thai sentence structure in the grammatical area. In addition, there are norms such as the use of Thai names of persons, places or objects, the retention of superstitious and redundant information and the supplementation of explanations for unfamiliar proper nouns or special words in the pragmatic area. Among the norms for the Thai MT, the use of vernacular Chinese, the SVO structure and the target languages for cultural terms can also be applied to western languages MT. However, the use of multisyllabic words, active voice and no use of makers to show the plural forms and past verb tense are different from those CC norms used for the MTs in western European languages, due to their different linguistic features as introduced in Section Two. Last, a questionnaire was done by ten respondents, and the findings clearly showed that the quality of the Thai MT was greatly improved when the CC texts were carefully controlled following the CC norms as identified above. The statistical results indicated that before the texts were controlled, the respondents' understanding was 42% on average, but that of CC texts rose up to 86.2% on average.

Even though this research has identified a set of norms exclusively designed for the Thai MT, there are still some limitations in the research, in terms of the small size of the text samples and the small number of respondents. As we all know, various topics can affect the style of writing. Consequently, writing CC texts for daily conversation and for the academic purpose could use different strategies and different norms. Moreover, the questionnaire was only done by ten Thai respondents, so more respondents should be invited to get more objective and more reliable results in the future research.

It is suggested that future researchers apply the similar research model to the MT application for the translation in different language pairs and for the translation of the texts of various genres, such as speech and fairy tales, to identify the similarities and differences in the CC norms. It is noted that different target languages have different linguistic attributes, so it is worth further investigating how they must be controlled to create the better quality MT outputs in diverse target languages, such as Japanese and Korean, Vietnam and Filipino and others.

References

Duanmu, S. (2007). *The phonology of standard Chinese*. UK: Oxford University Press.

Encyclopedia of Taiwan (2012). Religion & folklores [宗教 ・ 民俗]. Retrieved April. 25, 2013 from http://taiwanpedia.culture.tw/web/index

Jenks, P. (2006). On the Thai classifier-modifier construction. Retrieved May 12, 2013 from http://www.linguistics.berkeley.edu/~jenks/Research_files/HarvardTalk.pdf

Jenks, P. & Huang, S.Z. (2012). The functional architecture of nominal modifiers in Chinese and Thai. Retrieved May 15, 2013 from http://linguistics.berkeley.edu/~syntax-circle/Jenks-Huang_2012_booklet.pdf

Kullavanijaya, P. (2010). A study of some two-syllabled words in Thai. Retrieved April 30, 2013 from http://ebookbrowse.com/a-study-of-some-two-syllabled-words-in-thai-pdf-d396792772

Luo, S. & Sun, M. (2003). Two-character Chinese word extraction based on hybrid of internal and contextual measures. Retrieved from May 8, 2013 from http://acl.ldc.upenn.edu/W/W03/W03-1704.pdf

Mixdorff, H., Charnvivit, P. & Luksaneeyanawin, S. (2006). Realization and perception of tones in mono- and polysyllabic words in Thai. Retrieved from May 20, 2013 from http://public.beuth-hochschule.de/~mixdorff/thesis/files/mixdorff_charnvivit_issp2006.pdf

Potisuk, S. (2009). A lexicalized tree adjoining grammar for Thai. Retrieved May. 15, 2013 from http://www.aclweb.org/anthology-new/Y/Y09/Y09-2002.pdf

Schwitter, R., Ljungberg, A. & Hood D. (2003). A look-ahead editor for a controlled Language. Retrieved May. 8, 2013 from http://web.science.mq.edu.au/~rolfs/papers/CLAW03-ECOLE.pdf

Shih, C.-L. (2006). *Helpful assistance to translators: MT&TM*. Taipei: Bookman Books Ltd.

Shih, C.-L. (2011). *Real-time communication through machine-enabled translation:Taiwan's oracle poetry* [機器翻譯即時通：臺灣籤詩嘛ㄟ通].Taipei: Bookman Books Ltd.

Shih, C.-L. (2012). *Translation research models and application: Intra/Extra-linguistic perspectives.* Taipei: Bookman Books Ltd.

Sun Moon Lake National Scenic Area (n.d.). About Sun Moon Lake. Retrieved May.1, 2013 from http://www.sunmoonlake.gov.tw/AboutSunmoonLake/about01.htm

Yehliu Geopark (2009). Yehliu Natural Landscape. Retrieved May. 1, 2013 from http://www.ylgeopark.org.tw/content/landscape/Sight.aspx

Zhang, Y. (2006). Chinese word formation and terminology translation challenge. Retrieved from http://translorial.com/2006/12/01/chinese-word-formation-and-terminology-translation-challenge/