The Characteristics of Southeast Asian Languages and Their Influence on Translation

Lianfang LIU[1], Zixian DENG[1], Jiakai WEN[2], Liangchun LU[2], Yuanyuan PAN[3], Lixiang ZHAO[3]

(1. Guangxi Computing Center, Nanning, Guangxi 530022, P. R. China; 2. Guangxi Daring E-Commerce Services Co., Ltd., Nanning, Guangxi 530007, P. R. China)

Abstract: The exchange and cooperation between China and Southeast Asian countries serves as an important part of the Maritime Silk Road under the "Belt and Road" Initiative. Linguistic communication is a prerequisite for the exchange and cooperation. There are 9 major official languages, including English, Vietnamese, Thai, Malay, Indonesian, Lao, Burmese, Cambodian and Filipino in 10 countries of Southeast Asian. The latter 8 Southeast Asian languages are characterized by different grammars and are grammatically subsidiary to the Austroasiatic Language Family, Austronesian Language Family and Sino-Tibetan Language Family. The premise for works on the construction of Southeast Asian language resources and machine translation is to learn about these languages. As part of learning the basic knowledge of Southeast Asian languages, this paper introduces the characteristics of Southeast Asian languages and their influence on translation.

Keywords: Southeast Asian languages; Austroasiatic language family; Austronesian language family; Sino-Tibetan language family; translation

I. Terms and definitions

Analytic language：also known as “isolating language” or “radical language”, refers to a linguistic form characterized by: expressing grammatical relations by different word orders and function words when organizing sentences; using many compound words and few derivative words; using nouns without changes on gender, quantity and case.

Agglutinative language：refers to a linguistic form with rich morphological changes and expresses various grammatical relations via changes on forms of the words themselves. It is characterized by the combination of word roots and additional components and the overlap (or partial overlap) of word roots as the main means of word-formation and morpho-formation. The additional components are divided into prepositive, central and postpositive ones.

Loanwords：also known as "foreign words", refers to those words which are loaned from foreign languages phonetically and semantically. Every language has a certain number of loanwords. For example, the Chinese words like “葡萄”, “石榴”, “狮子” and “玻璃” were loaned from the Western Region in the Han Dynasty of ancient China; the Chinese Buddhist terms like “佛”,“菩萨”, “罗汉”, “尼”, “和尚” were borrowed from the ancient India after the Han Dynasty; and words like “胡同” and “站” were loaned from the ancient Mongolia during the Chinese Yuan Dynasty.

Tonal language：refers to languages in which meanings of single words are distinguished with tones, and is characterized that different meanings are formed by tones of different lengths and levels without changes in pronunciation.

II. Linguistic Characteristics

The 8 Southeast Asian languages are subsidiary to the Austronesian Language Family, Austroasiatic Language Family and Sino-Tibetan Language Family respectively, present a large number of common characteristics yet have their own features.

2.1 In Morphology

The Austroasiatic Vietnamese and Cambodian and the Sino-Tibetan Thai, Burmese and Lao are analytic languages just like Chinese. In these languages, the grammatical relations are expressed mainly by word orders and function words without changes on gender, quantity and case of nouns.

The Austronesian Indonesian, Malay and Filipino are agglutinative languages with rich morphological variations which can express various grammatical relations. The combination of word roots and additional components (i.e. prefix, infix and suffix) and the overlap (or partial overlap) of word roots is the main means of word-formation and morpho-formation.

2.2 In Phonetics and Tones

The Austroasiatic Vietnamese and the Sino-Tibetan Thai, Burmese and Lao are tonal languages like Chinese. In these languages, different meanings are formed by tones of different lengths and levels without changes in pronunciation. Vietnamese and Lao thereof have the richest tones of up to six.

The Austroasiatic Cambodian and the Austronesian Indonesian, Malay and Filipino belong to non-tonal languages (namely "intonation languages") like English, where the phonetic tones in different lengths represent only tones rather than meanings.

2.3 In Writing System

The Austroasiatic Vietnamese and the Austronesian Indonesian, Malay and Filipino adopt Latin alphabet. The spelling of Vietnamese is even more complicated, including 7 Latin letter variants (letters for Vietnamese only) and 6 tone symbols, where careless misspelling would lead to wrong meanings.

The Austroasiatic Cambodian and the Sino-Tibetan Thai, Burmese and Lao adopt their own unique but similar-sourcing alphabet letters.
These languages are quite different from the pictographic and ideographic Chinese in writing.

2.4 In Syntax

Vietnamese, Cambodian, Indonesian, Malay, Thai and Lao in the three major language families all use the same basic order of "Subject-Predicate-Object" as Chinese, yet what is different from Chinese is that the modifiers are placed after the central words modified.

Moreover, the Austronesian Filipino uses quite unusual word orders like

“Object - Predicate - Subject” or “Predicate - Subject - Object”; and the Sino-Tibetan Burmese adopts a word order of“Subject - Predicate - Object” but the modifiers are placed before the central words like Chinese.

2.5 In Vocabulary

All of these languages contain a certain amount of loanwords (i.e. "foreign words"), and have been affected to varying degrees by Sanskrit and Pali since the introduction of Buddhism.

Cambodian, Malay, Thai, Lao and Burmese are most affected by Sanskrit. Most loanwords from Sanskrit and Pali languages are polysyllabic words and still retain the original gender and quantity characteristics. Vietnamese is deeply influenced by Chinese, French and English, specifically, loanwords from Chinese account for about 60% of all the loanwords in Vietnamese and even up to 70-80% in political, economic and legal fields while technical words are mainly loaned from French and English. Indonesian is heavily influenced by the Dutch system and contains a large number of Javanese and Dutch loanwords in its vocabulary. Filipino vocabulary is deeply affected by Spanish.

In addition, Malay and Indonesian differ slightly in the pronunciation and vocabulary of the writing system, and people who use these two languages are basically able to communicate with each other. The Javanese and Dutch loanwords in Indonesian are the main reason for this difference.

III. Difficulties in human translation

It is generally known that the development and maturity of an industry requires the accumulation of time and practice. However, the translation industry for Southeast Asian languages does not have the two conditions in nature comparing with generally used languages like English.

3.1 Transfer between syntactical meanings

This is one of the major difficulties for beginners in translation for Southeast Asian languages. Especially when translating complex long sentences, if not

handle in a proper way, they may produce "translationese" (foreign-styled Chinese or Chinese-styled foreign language) that would be neither precise nor fluent and even result in serious mistranslations that are completely contrary to the original meanings.

Taking Chinese to Thai translation for instance, there are often many gorgeous attributes appearing before central words in Chinese sentences which make it difficult for translator to identify these attributes, to decide how to put the translation after central words in Thai without producing literal faults and to find enough gorgeous Thai words.

Taking Indonesian to Chinese translation as another example, the "Yang" structure serves as an important linkage in long sentences of Indonesian. It is often difficult for translators to read and translate such a long sentence which describes many things without using any punctuation and with multi-level clauses. Any mistake would lead to literal faults of the entire sentence or even the paragraph.

3.2 Translation of loanwords

Vocabularies of Cambodian, Thai, Lao and Burmese contain a large number of Sanskrit and Pali loanwords used in religious and aristocratic life. It is very difficult to memorize and use these words. For one thing, Sanskrit is considered as the world's most difficult language to learn for its polysyllable words and changes on gender and quantity. For another thing, due to limitation in cultural and developing level of Southeast Asian countries, English loanwords are commonly used and written in the countries' own special writing system to express modern vocabulary. Thus it is difficult for translators who are not familiar with English to translate Southeast Asian languages into Chinese quickly.

The Chinese loanwords in Vietnamese can be divided into 3 categories: phonemical & semantical, phonemical only, and Vietnamized. Translators should distinguish them and make good understanding to translate them correctly.

Indonesian vocabulary contains a number of Dutch and Javanese loanwords that cannot be found in dictionaries. In addition, English loanwords spelled in Indonesian pronunciation are commonly used probably because they use the same Latin alphabet writing system. For example, some project contracts and documents for bidding and tendering (accounting for about 70% of Indonesian documents needed to be translated in the market) often contain Indonesian-styled English words that cannot be found in dictionaries and are difficult to be translated. Conversely, translators would find it hard to complete their work on translating Chinese engineering and mechanical documents into Indonesian if they are not familiar with English and unable to produce "Indonesian-styled" English loanwords.

3.3 Conversion of Punctuations

In countries colonized by French or Dutch in the history like Vietnam and Indonesia, a comma (",") is often used as a decimal point and a full stop (".") is used as a thousand separator. However, the Chinese usage is quite the contrary. For example, "12,345" in Vietnamese should be translated into "12.345" in Chinese and vice versa.

In Thai language, there are no comma, full stop, question mark, exclamation and other punctuations. The text takes the form of continuous writing without any punctuation and space between words, and uses an interval of two letters or a small pause in sentences to indicate the ending of a sentence. In that case, the translators should analyze the context carefully to judge the boundaries of words and sentences and decide how to do the translation. It is easy to make mistakes if translators have poor vocabulary or contextual thinking.

IV. Difficulties in Machine Translation

4.1 Text Encoding

Comparing with languages based on Latin alphabet writing system such as Indonesian, Malaysian, and Filipino, it is more difficult for other Southeast Asian languages in text recording, recognizing and handling. On one hand, the data entry can only be completed by a person who knows how to type the language of the country with specified fonts, or else the typing results cannot

be recognized by the system. However, as for Indonesian, anyone who can type in 26 letters can enter data and check the texts. On the other hand, there are few or even no available tools to support text encoding of these languages, which requires specific research and development work. Besides, it is easy to generate messy codes during switching and processing of different procedures and tools.

4.2 Syntactical Translation

The major difficulties for Rule-based Machine Translation (RBMT) lie in the differences of syntactic rules between two languages and the analysis of sentence constituents.

Despite the current popular Neural Machine Translation (NMT) may not have to face the difficulties mentioned above, its features were limited greatly by lacking of corpora.

4.3 Segmentation Rules

In Thai, there is no any ending punctuation and clear segmentation rules, which impacts the results of Machine Translation to a certain extent. In Burmese, full stops are written in special Burmese character, which should be considered in Machine Translation.

4.4 Translation of Loanwords

Most loanwords are unlogged words without corresponding meanings in corpora, and they require human translation.

4.5 Conversion of Punctuation

As same as human translation, the usage of decimal points and digital symbols in Vietnamese and Indonesian is quite contrary to that in Chinese. How to judge and convert punctuations correctly and automatically to avoid serious quality defects remains an issue that needs to be solved.

References:

[1] WANG Hui. The Language Situation and Language Policy in the “Belt and Road” Countries, Volume 1. SOCIAL SCIENCES ACADEMIC PRESS (CHINA).

[2] LIN Minghua. A Brief Discussion on written Vietnamese [J]. Modern Foreign Languges, 1983(3):55-59.

[3] TAN Zhici. A Primary Analysis on Reasons for which Written Vietnamese is Profoundly Influenced by Chinese Characters [J]. SOUTHEAST ASIAN AND SOUTH ASIAN STUDIES, 1998(2):47-50.

[4] CHEN Hui. The Situation and Developing Trend of Language Department of South-east Asia in China[J]. 2007(3):72-75.