Identifying Idioms in Chinese Translations

Wan Yu Ho, Christine Kng,* Shan Wang and Francis Bond
Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore
* **St. John's College**, Santa Fe; **Hopkins-Nanjing Center**, Nanjing
L120002@e.ntu.edu.sg, kngcyl@gmail.com, wangshanstar@gmail.com, bond@ieee.org

Abstract

Optimally, a translated text should preserve information while maintaining the writing style of the original. When this is not possible, as is often the case with figurative speech, a common practice is to simplify and make explicit the implications. However, in our investigations of translations from English to another language, English-to-Chinese texts were often found to include idiomatic expressions (usually in the form of Chengyu 成语) where there were originally no idiomatic, metaphorical, or even figurative expressions. We have created an initial small lexicon of Chengyu, with which we can use to find all occurrences of Chengyu in a given corpus, and will continue to expand the database. By examining the rates and patterns of occurrence across four genres in the NTU Multilingual Corpus, a resource may be created to aid machine translation or, going further, predict Chinese translational trends in any given genre.

Keywords: Translation shift, Idioms, Chinese, English, Chengyu

1. Introduction

Due to intrinsic syntactic differences, as well as culture-influenced semantic differences between languages (or language dialects), it can be difficult or even impossible to achieve a perfect translation in which all information is transmitted from one language to another. In most cases, one has to decide the parts of information to be left out so as to achieve the best possible translation, as it is often not possible to preserve the exact style of writing, its evoked senses, or both, in translation.

Baker (1992, 2007) presents four translational universals: explicitation, standardization, simplification, and normalization, and there appears to be overlaps between explicitation and simplification, and standardization and normalization.

Explicitation states that when it is not possible to preserve both meaning and style, it is preferable to make explicit the implied meanings instead, sacrificing style for semantic fidelity.

Simplification refers to a "reduction in lexical density" (Ghadessy, M. & Gao, Y., 2001), wherein the translation tends to feature many more frequently-used or common words as compared to the original text.

In standardization, the translated text tends to trend towards a more formal register or the standard grammar of the target language, while normalization refers to a tendency of translators to follow the conventional writing styles of the target language and stay away from idiosyncratic or otherwise creative uses of the language.

1.1 Translating Idioms

An idiom is a (usually fixed) expression "whose meaning cannot be predicted from the meanings of the constituent words" (Collins English Dictionary, n.d.). As idioms evoke additional senses to the figurative meaning, they are also often indicative of and encapsulate the culture in which they originate. While they "enhance naturalness and create an impression of fluency" (Baker, 2007), their cultural specificity also means that it is often nearly impossible to fully translate all of the evoked senses into the target language.

For cases in which an exact idiomatic equivalent or approximation cannot be found, a translator would normally replace the idiom with a non-idiomatic synonym, for it is more important to preserve the idiomatic meaning rather than the evoked senses.

If translating an idiom into a non-idiomatic phrase results in the shedding of information and meaning, then the opposite would also be true: translating a non-idiom to an idiomatic expression would add information and meaning to the text. Nida and Taber (1974) explain this as allowing the message to "speak meaningfully to people in terms of their own lives and behaviour". Perhaps the inclusion of culturally-specific expressions enables the reader to better understand a piece of writing by framing it in a context which they would be intimately familiar with.

However, the general preference of translators to translate idioms into cultural equivalents, approximants, or non-idioms instead of the other way around would mark the latter as an exception.

We would hence expect English-to-Chinese translations to be simplified and made explicit, as is par for the course. However, while tagging concepts across languages in the NTU Multilingual Corpus (Tan & Bond, 2011), or NTU-MC, it was noted that, contrary to predictions, idiomatic expressions often appeared in the translated Chinese texts where there was none in the original (Kng & Bond, 2012).

To investigate this further, we have created a small lexicon of Chengyu, using which we tagged all occurrences of Chengyu across the four genres in the NTU-MC.

We hope that by examining frequency patterns of Chengyu in translations into English-to-Chinese translated texts of different genres, a resource can be created to aid machine translation, and perhaps predict translational trends in Chinese translations based on a given genre, including the most common Chengyu used.

2. The NTU Multilingual Corpus

The NTU-MC was created and made public in 2011 (Tan & Bond, 2011) was still developing. Currently it comprises approximately 595,000 words (26,000 sentences) in seven languages (Arabic, Chinese, English, Indonesian, Japanese, Korean and Vietnamese) from seven language families (Afro-Asiatic, Sino-Tibetan, Indo-European, Austronesian, Japonic, Korean as a language isolate and Austro-Asiatic) (Tan & Bond 2012; Bond et al. 2013). This allows for the comparison of the same text in different languages, as well as aids investigations into translational trends between languages.

The parallel multilingual texts have been expanded to comprise several different genres (*story*, *news*, *essay*, and *tourism*) compiled from a variety of sources. The *story* genre has the two *Sherlock Holmes* short stories *The Adventure of the Speckled Band* and *The Adventure of the Dancing Men*. The *news* genre comprised publications from Kyoto University Corpus[1] (Kurohashi and Nagao, 2003). The text used for the *essay* genre was *The Cathedral and the Bazaar* (Raymond, 1999); the *tourism* genre comprised text from the Official Singapore Tourism Website[2].

[1] http://nlp.ist.i.kyoto-u.acp.jp
[2] www.yoursingapore.com

3. Chengyu (成语)

Chengyu, often translated into English as "Chinese idioms" are prototypically four-character, non-compositional phrases derived from historical lore or classical literature. Wu (1995, pp. 81) describes Chengyu as follows:

> "The Chinese idiom 'Chengyu ' is a set phrase, an old expression, prevalent in society, used by the common folk, has seen ages of constant use, usually in four-character form with varying constituent constructions and diverse origins. The meanings for some of the idioms can be deduced from their composite constituents. By contrast, with some of them, their meanings cannot be gained from their constituents unless we know the semantic fields or historical sources. The fixed form in its structure and semantics is its critical characteristic. It functions as a lexeme in sentences and behaves more vividly and symbolically than its synonyms represented by common lexemes. Its formation can be derived, inherited, or borrowed."

A Chengyu is similar to the English idiom in that it is a 'frozen' expression and expresses a meaning not necessarily derivable from its constituents. However, as most, if not all, Chengyu are derived from historical lore, classical literature, or Chinese culture, Chengyu also frequently evoke those sources to add a further layer of symbolic meaning to the text, in addition to its compositional or literal meaning.

(1) shows an example of a prototypical Chengyu:

守	株	待	兔
shǒu	*zhū*	*dài*	*tù*
to guard	tree stump	wait	rabbit

LIT: waiting by a tree stump for a rabbit
'to expect fortune without putting in effort'

(1) A prototypical Chengyu

Furthermore, unlike English idioms, it is also possible to have Chengyu which are fully compositional and have no figurative meaning. In those cases, their status as Chengyu come from their historical or literary heritage.

An example of a Chengyu with no figurative meaning is shown in (2):

分	崩	离	析
fēn	*bēng*	*lí*	*xī*
divide	rupture	leave	split apart

'to completely fall apart'

(2) A Chengyu with only compositional meaning

分崩离析 has no extended metaphorical meaning, but its origins in classical literature, namely the Analects (Confucius, trans. 1861), determine its status as a Chengyu.

Some Chengyu have their source in idioms or metaphors, though a vast majority appears to originate from ordinary culture, such as folk tales, famous texts of ancient times, or even simply everyday speech, while others arose from the influences of foreign culture, such as Buddhism.

The use of Chengyu is usually regarded as not only a sign of eruditeness, but they also contribute a pleasing rhythm to reading, and the multiple evoked senses help to keep the text interesting and full of flavour. This, combined with the ability of Chengyu to appear in several different parts of speech, could explain the prevalence of Chengyu in Chinese texts and translations, as it would allow Chengyu to be readily inserted without affecting the overall tone or register as idioms might in English texts.

3.1 Other Idiomatic Expressions in Chinese

In addition to Chengyu, there also exist several other forms of idiomatic expressions, such as: Guanyongyu 惯用语, Xiehouyu 歇后语, Yanyu 谚语, Geyan 格言, Jingju 警句, Suyu 俗语, Liyu 俚语.

Not all are necessarily four-character phrases like Chengyu; some, such as Suyu, may take the form of phrases or short sentences:

身	在	曹	营		心	在	汉
shēn	*zài*	*cáo*	*yíng*		*xīn*	*zài*	*hàn*
body	at	Cao	encampment		heart	at	Han

LIT: to have one's body in Cao Cao's encampment, but one's heart with the Han people
'to not have one's mind on one's work; be distracted'

(3) A Suyu

Although there is currently no fixed consensus on their membership statuses, those idiomatic expressions may be considered subsets of the general term Shuyu 熟语, which is believed to describe an idealised idiomatic expression instead of having a fixed definition (Huang & Liao, 2002).

Shuyu hence refer more to the shared characteristics of its subsets, which Huang (2007) states to be: (1) a fixed structure, with (2) a fixed idiomatic meaning; (3) has been in frequent use since historical times; (4) have specific places for pauses when reading, and (5) rhythmic symmetry between the two pause-segmented subunits.

Furthermore, the use of Shuyu, likely due to its connections to historical culture and literature, can not only convey a country's culture (namely, China, or Chinese culture), but it also displays one's standard of the language.

Although Shuyu and other idiomatic expressions are not part of the current focus, we hope to eventually incorporate them into our lexicon in the future.

4. The Chengyu Database

An initial list of about 4, 000 Chengyu was created by combining several lists available online: Wiktionary (Wikimedia, 2013), online Chinese-English (chinesenotes.com, 2011) and Japanese-English dictionaries (Breen, 1991), and four-character words in the *Academica Sinica* corpus distributed with the Natural Language Toolkit (Bird et al, 2007). Items which were not Chengyu were manually removed.

As Chinese does not inflect, and Chengyu are usually 'frozen' expressions, we can pick out all instances of listed Chengyu in a text using simple string matching.

The initial run of the list, using *The Cathedral and the Bazaar*, showed only 11 Chengyu out of a manually-identified 35, indicating that the list could only identify about 31% of the Chengyu present. To rectify this, we went through the corpus manually to pick out missed Chengyu and remove four-character phrases which had been mistakenly listed as Chengyu. We are not trying to tailor our Chengyu list to fit the NTU-MC, but rather using it as another starting source of data to expand the database.

As the addition (or removal) of listed items is currently a largely manual process, it is possible that mistakes will be made in modifying the list at some point, resulting in Chengyu being left out or non-Chengyu being added. However, continuous cycles of string matching, manual reviews and modifications should improve the list's accuracy while steadily expanding the lexicon with each revision. Our final goal is to integrate this list with the Chinese Open Wordnet (Bond & Foster, 2013; Wang & Bond, 2013a, 2013b).

Each full entry will be a new synset, with, minimally: an index form, a Chinese definition, an English definition (possible with separate literal and figurative meanings), a *domain-usage* link to the entry for Chengyu, semantic links to existing entries, and any examples found in our corpora (with genre and sentence id). A further possibility is a link to other lexicons, such as dictionaries in which each entry may be found. An example entry is (4):

Index	根深蒂固	
Definition	English	deep-rooted (LIT: roots deep stem strong)
	Chinese	基础深厚，难以动摇
Link	domain-usage	***chengyu***$_{n:1}$
	similar-to	***deep-rooted***$_{n:1}$
Example	essay:176	罪魁祸首自然是那些根深蒂固的错误和持续的恶性循环。 *In the cathedral-builder view of programming, bugs and development problems are tricky, insidious, deep phenomena.*
	news: 101313	另外，对再次被日本统治的警戒论也根深蒂固。 *Furthermore, the Japanese conquest had left a deep, lasting trauma.*
	news: 101421	对贝鲁斯柯尼的搜查是成为促使这个战后政治崩溃之契机的净化作战的一环，问题根深蒂固。 *An investigation into Berlusconi has become an opportunity to clear away the post-war political competition, a deep-rooted problem.*

(4) An entry in the Chengyu database

5. The Annotated Corpus

Looking at our current corpus, Chengyu have a rather high rate of occurrence as shown in Table 1, particularly in the *essay* genre where a Chengyu occurs once every 6 or 7 sentences. The *story* genre has the highest percentage of "types" of Chengyu while other genres have more repeated ones. This may be an evidence that stories use more Chengyu to better evoke imagery.

A preliminary investigation of a few of the most frequently-occurring Chengyu also reveals some pattern regarding the meaning of each Chengyu used.

This is particularly evident in the t*ourism* genre, where the most frequently-used Chengyu (of which 23 are shown in Table 2) revolved around sensory imagery (such as 大快朵颐 and 琳琅满目) or hyperbolic expressions of variety or quality.

This is not unexpected, as Chengyu have fixed, distinct meanings, and would hence only be suitable in particular contexts. In the *tourism* genre, as the aim is to entice readers to visit and experience the country, the writer or translator would hence focus only on Chengyu which evoke the five traditional senses, restricting the range of Chengyu available for the writer to use. This could also explain why the percentage of *unique*, or types of Chengyu seem unusually low (48.2%) as compared to the other three genres.

If such patterns persist as the corpus grows, we may be able to use this data to predict not only the rates of Chengyu occurrence in a text but also the kinds of Chengyu that will be used, based on a given genre.

However, it must be noted that as our corpus is still relatively small, our current numbers may have been affected by the idiosyncrasies of the translators or texts, affecting what we would perceive as a general trend of Chinese translations.

6. Discussion and Future Work

There are many Chinese idiom dictionaries, with some containing over 20,000 Chengyu, often containing many literary examples and commentary on the source of the idiom (Jiao et al., 2011). Given the long heritage of Chengyu, it is possible that the actual number of published dictionaries is much higher.

Instead of duplicating this work on a digital platform, our goal is to produce a compact lexicon, available for natural language processing, which can be integrated with existing lexical resources to aid ongoing work on studying translation shift (Bond et al., 2013).

Genre	Story	News	Essay	Tourism	Overall
Sentence/#	1, 226	2, 138	816	3, 280	7,460
Chengyu (token)/#	90	161	127	388	766
Chengyu (type)/#	79 (87.8%)	108 (67.1%)	103 (81.1%)	187 (48.2%)	427 (55.7%)
Chengyu per 100 Sentences	7.3	7.5	15.6	11.8	10.3

Table 1: Distribution of Chengyu in four different genres

Genre	Chengyu	Pinyin	English Translation	Count
Story	无论如何	*wú lùn rú hé*	"no matter the circumstances"	3
News	全力以赴	*quán lì yǐ fù*	"to go all-out; spare no effort"	3
	动荡不安	*dòng dàng bù ān*	"in turmoil"	3
	堆积如山	*duī jī rú shān*	"to pile up, like mountains"	3
	无论如何	*wú lùn rú hé*	"no matter the circumstances"	3
	迫在眉睫	*pò zài méi jié*	"imminent"	3
	各种各样	*gè zhǒng gè yàng*	"a wide variety"	5
	竭尽全力	*jié jìn quán lì*	"to use all of one's strength or resources"	5
	引人注目	*yǐn rén zhù mù*	"to attract a lot of attention; highly conspicuous"	7
	下落不明	*xià luò bù míng*	"to have one's whereabouts be unknown"	8
	理所当然	*lǐ suǒ dāng rán*	"an assumed certainty"	10
Essay	半途而废	*bàn tú ér fèi*	"to give up halfway, wasting previous effort"	3
	显而易见	*xiǎn ér yì jiàn*	"obvious, evident"	5
Tourism	古色古香	*gǔ sè gǔ xiāng*	"antique flavours"	3
	耳目一新	*ěr mù yī xīn*	"refreshing"	3
	顾名思义	*gù míng sī yì*	"as the name implies; self-explanatory"	3
	五花八门	*wǔ huā bā mén*	"a wide, kaleidoscopic variety"	3
	各行各业	*gè háng gè yè*	"a variety of occupations"	4
	熙熙攘攘	*xī xī rǎng rǎng*	"bustling with activity"	4
	目不暇接	*mù bù xiá jiē*	"to have so many details, they cannot be seen all at once"	4
	闻名遐迩	*wén míng xiá ěr*	"extremely well-known; world-renowned"	4
	叹为观止	*tàn wéi guān zhǐ*	"breathtakingly impressive"	5
	大街小巷	*dà jiē xiǎo xiàng*	"every nook and cranny of every street"	5
	相映成趣	*xiāng yìng chéng qù*	"to contrast, usually in a complementary manner"	5
	不胜枚举	*bù shèng méi jǔ*	"too many to count"	6
	前所未有	*qián suǒ wèi yǒu*	"unprecedented"	6
	垂涎欲滴	*chuí xián yù dī*	"to desire something so much, one is drooling"	6
	讨价还价	*tǎo jià huán jià*	"bargaining"	6
	首屈一指	*shǒu qū yī zhǐ*	"to be second to none"	6
	各式各样	*gè shì gè yàng*	"all types and kinds"	7
	大快朵颐	*dà kuài duǒ yí*	"to heartily enjoy a meal"	8
	琳琅满目	*lín láng mǎn mù*	"a feast for the eyes"	9
	流连忘返	*liú lián wàng fǎn*	"to be so utterly captivated that one forgets about home"	10
	独一无二	*dú yī wú èr*	"unique; without compare"	10
	应有尽有	*yīng yǒu jìn yǒu*	"to have everything one might expect or desire"	14
	各种各样	*gè zhǒng gè yàng*	"a wide variety"	18

Table 2: Chengyu occurring at least 3 times in each genre

We will release the Chengyu list and other additions to the Chinese Open Wordnet[3] under the same license as the Princeton Wordnet (Fellbaum, 1998). The tagged corpora will all be released as part of the NTU-MC[4] (CC-BY).

The license allows the free adaptation and distribution of the work, as long as the original source is attributed.

In addition to expanding the Chengyu database, we will also eventually return to our original problem and investigate Chengyu both monolingually and in parallel with the source text(s).

We would also like to look at Chengyu or Chengyu-like expressions in other languages such as Korean and Japanese, where they are also used, but less often than in Chinese.

[3] http://compling.hss.ntu.edu.sg/cow
[4] http://compling.hss.ntu.edu.sg/ntumc

7. Acknowledgements

We would like to thank Ning Chuang-Goodman for help in creating the Chengyu lexicon. We wish to acknowledge the funding for this project from Nanyang Technological University under the Undergraduate Research Experience on CAmpus (URECA) programme and the MOE Tier 1 grant *Shifted in Translation—An Empirical Study of Meaning Change across Languages* (RG51/12). This research was started when the second author visited NTU as an Ariel Intern from St John's.

8. References

Baker, M. (1992). *In Other Words: A Coursebook on Translation*. Routledge.

Baker, M. (2007). Patterns of Idiomaticity in Translated vs. Non-Translated Text. *Belgian Journal of Linguistics,* 21(1):11-21.

Bird, S. (2001). *Natural Language Toolkit, Sinica Treebank Reader.* Retrieved 14 March, 2014, from http://nltk.org/_modules/nltk/corpus/reader/sinica_treebank.html.

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, pp.1352-1362.

Bond, F., Wang, S., Gao, E. H., Mok, H. S, & Tan, J. Y. (2013). Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013),* Sofia, pp. 149–158.

Breen, J. (1991). *Japanese/English Dictionary Project*. Retrieved 14 March, 2014, from http://www.csse.monash.edu.au/~jwb/edict_doc.html .

Chinesenotes.com. (2011). *Chinese-English Dictionary*. Retrieved 14 March, 2014, from http://chinesenotes.com/dictionary_reuse.php.

idiom. (n.d.). *Collins English Dictionary - Complete & Unabridged 10th Edition.* Retrieved March 19, 2014, from Dictionary.com website: http://dictionary.reference.com/browse/idiom

Confucius, 1861. 論語[*The Analects*] (Legge, J., Trans.). Available from the Chinese Text Project, at: http://ctext.org/analects .

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Huang, B. R. & Liao, X. D. (2002). *Modern Chinese, 3rd Ed., Volume 1. [现代汉语（增订三版）（上册）].* Beijing: Higher Education Press (高等教育出版社).

Kurohashi, S. & Nagao, M. (2003). Building a Japanese parsed corpus — while improving the parsing system. In Anne Abeillé (ed.) Treebanks: Building and Using Parsed Corpora, Chapter 14, pages 249–260. Kluwer Academic Publishers.

Jiao, L., Kubler, C. C., & Zhang, W. (2011). *500 Common Chinese Idioms: An annotated Frequency Dictionary*. Routledge.

Raymond, E. S. (1999.) The Cathedral & the Bazaar. O'Reilly.

Tan, L. & Bond, F. (2011). Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation* (PACLIC 25), Singapore, pp. 367–376.

Tan, L. & Bond, F. (2012). Building and annotating the linguistically Diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing 22 (4),* pp. 161–174.

Wang, S. & Bond, F. (2013a). Building the Chinese Open Wordnet (COW): Starting from Core Synsets. In *Proceedings of The 11th Workshop on Asian Language Resources, Workshop of The 6th International Joint Conference on Natural Language Processing (IJCNLP-6)*, Nagoya, Japan, pp.10-18.

Wang, S. & Bond, F. (2013b). Theoretical and Practical Issues in Creating Chinese Open Wordnet (COW). Paper presented at *The 7th International Conference on Contemporary Chinese Grammar (ICCCG-7)*, Nanyang Technological University, Singapore.

Wikimedia Foundation, Inc. (2012). *Wiktionary*. Retrieved 14 March, 2014, from http://zh.wiktionary.org/wiki/Category: 成語.

Wu, C. (1995). On the Cultural Traits of Chinese Idioms. *Intercultural Communication Studies, 1, 61-84.*