Bilingual Multi-word Expressions, Multiple-correspondence, and their Cultivation from Parallel Patents: The Chinese-English Case

Benjamin K. Tsou
City University of Hong Kong
The Hong Kong University of Science and Technology / Hong Kong SAR
btsou99@gmail.com

Ka Po Chow
Chilin (HK) Ltd.
Hong Kong SAR
Kapo.rclis@gmail.com

John Lee
City University of Hong Kong
Hong Kong SAR
jsylee@cityu.edu.hk

Ka-Fai Yip
Yale University / New Haven
Connecticut, United States
kafai.yip@yale.edu

Yaxuan Ji
The Hong Kong University of Science and Technology / Hong Kong SAR
yjiaf@connect.ust.hk

Kevin Wu
Chilin (HK) Ltd.
Hong Kong SAR
kjwuhk@gmail.com

Abstract

Multi-Word Expressions (MWEs) typically offer challenges in both linguistics and Natural Language Processing (NLP) and their cross-lingual correspondences also introduce new issues. This paper draws on a specially cultivated corpus of more than 300,000 comparable Chinese-English patents over 10 years [Patentlex: http://patentlex.chilin.hk], and focuses on issues related to bilingual correspondence between Chinese and English technical vocabularies extracted from it in terms of: (1) Non-unique correspondence between cross-lingual terms, which so far has not attracted sufficient interests, (2) Means to cultivate good sources for up-to-date technical terms, (3) A network approach to the weighted multilingual alternate renditions and their presentation through knowledge graphs, and (4) Typological differences in the cross-lingual MWEs, including the internal structure of constituent words and their sociolinguistic-discoursal registers.

1. Introduction

Compound words usually contain more than one constituent words (e.g. *watering hole, space station*) and multi-word expressions (MWEs) contain two or more constituent words (e.g. *sodium bicarbonate, subdural hematoma, ASAP, kicking the bucket, still water runs deep* [1]). MWEs are sometimes referred to as phrasal words, and they can be quasi-autonomous constructions within a sentence. Some have *locus classicus* (e.g. "probing for his Achilles' Heel" [2]) and are within the repertoire of only the well-educated or of those in technical and specialized fields. MWEs typically provide learning challenges for non-native speakers of the language as well (Foster et al. 2014, Wray 2002). The use of MWEs in language is quite pervasive (Jackendoff 1997), approaching half of the adult lexicon (Sag et al. 2002) with reference to WordNet (Fellbaum 1998).

The search for more sophisticated and more extensive resources involving MWEs has surged forward following the accelerated developments in science and technology in the run up to the new Millennium, and with the dramatic improvements in computer power, machine learning and AI to handle big data. As shown in

[1] Idiomatic expressions are also examples of MWEs, and they are found in abundance in many Asian languages (Tsou 2012).

[2] This refers to the weakest part of Achilles's body and is a metaphorical reference to an individual's weakness.

Figure 1, patent registration has seen phenomenal growth in recent decades with China taking the lead since 2016.

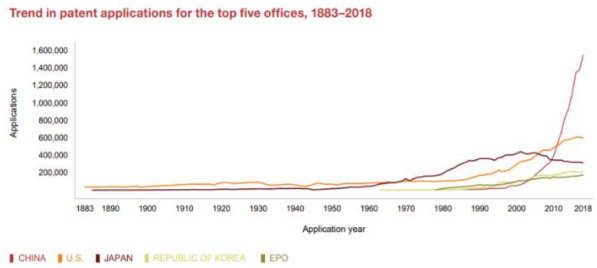

Figure 1. Trend in patent applications

Following rapid changes such as the above and in view of evolving global trade and international relations, there is increasing recognition of the need to overcome cross-lingual communication gaps. Effective efforts to do so entail the processing of MWEs, including handling them in Chinese/English cross-lingual NLP which involves complex texts such as technical manuals, legal documents, contracts and patents in NMT, cross-lingual retrieval and related data analytics.

This paper is organized as follows. Sect. 2 addresses the problems of multiple renditions for MWE translation. Sect. 3 proposes a possible solution, Multiple-Rendition Index, which requires a rigorously cultivated corpus. Sect. 4 introduces our corpus Patentlex and its data cultivation and curation. Sect. 5 compares MWEs in Patentlex with other corpora. Sect. 6 discusses typological differences in the cross-lingual MWEs in terms of correspondence, network representations, stratification and word structure. Sect. 7 concludes.

2. Problems of multiple renditions

There is no simple and regular one-to-one cross-lingual correspondence between any lexical pair of languages, except for very close relatives of the same dialect. For example, the corresponding terms for English mono-morphemic words such as *beef*, *mutton*, and *pork* are respectively bi-morphemic words in Chinese: *niurou* 牛肉, *yangrou* 羊肉, *zhurou* 豬肉 as well as *tunrou* 豚肉. A bilingual dictionary should have the corresponding terms or otherwise inappropriate translations such as the following may result: "I will eat *cow-meat* (for beef), *sheep-meat* (mutton), and *pig-meat* (pork)". We could also have *small cow meat* for veal and either *small lamb* (*xiaoyang* 小羊) or *small lamb meat* (*xiaoyangrou* 小羊肉) for lamb, where the differences between eating meat of the small lamb and eating the small animal could be significant. A single-word or multi-word term may have different renditions in translation according to different contexts. The non-one-to-one correspondence between the simpler kinship terms in Western languages, for example, and the more complex ones in Asian languages are also good illustrations of cultural and cognitive differences being foregrounded. This is especially so for technical terms found in scientific and technical texts and typified by patents. For example, the word "*multiplication*" has more than one word sense.

"Multiplication①: *cheng* 乘"

English: In particular, the invention relates to performing dual complex **multiplication** and complex division using a common circuit.

Chinese Translation: 尤其是, 本发明与使用一共同电路执行双复数乘法及复数除法有关。[3]

"Multiplication②: *fanzhi* 繁殖"

English: Growth and **multiplication** of microbes is substantial when it changes the viscosity, stability, or other important property of the composition.

Chinese Translation: 当微生物的生长与繁殖改变了该组合物的粘性、稳定性或其他重要特性时，微生物的生长与繁殖是实质性的(substantial)。[4]

These examples show that multiplication has a precise meaning in terms of the number of replications [*Cheng* 乘, a simple word in Chinese] in mathematics or physics, as in the first example. It can have an imprecise meaning [*Yansheng* 衍生 "derive-generate", a compound word in Chinese] when referring to biological reproduction in the second example. There are likewise different senses of "base" in botany, chemistry and electronics.[5]

2.1 Value of authentic usage statistics

The following is a good illustration of a common situation when an MWE with multiple renditions is encountered in the translation of technical texts:

"The invention relates to a steam jet enthalpy heat pump air-conditioning hot water unit which at least comprises a compressor, a

[3] According to International Patent Classification (IPC), this word sense is found mostly in domain G (Physics).

[4] Mostly found in domain A (Human Necessities) in patents.

[5] The development of technical terms in Chinese has a relatively short history, see Shen (2001) and Amelung (2004).

four-way reversing valve, the four-way reversing valve, an outdoor heat exchanger……"

There could be several MWEs within the single technical term found in the above passage which may not be familiar to a translator: (1) *steam jet*, (2) *enthalpy*, (3) *heat pump*, (4) *air-conditioning hot water unit*. If s/he is able to look up an appropriate dictionary, s/he may be bewildered by the multiple alternate renditions available and may have difficulty in obtaining useful information to facilitate any translation task at hand. For each constituent term there are multiple renditions available, as can be seen in the top three alternate renditions given in each case in the following examples. It should be clear that among the examples there may not be easy means to decide on an appropriate choice.

(1) Steam jet:
a. 蒸汽喷射(71.42%)
b. 蒸汽射流(21.42%)
c. 蒸汽喷射器(7.14%)

(2) Heat pump
a. 热泵(98.19%)*
b. 加热泵(0.96%)
c. 供热泵(0.69%)

(3) Enthalpy
a. 焓(88.43%)
b. 焓变(6.02%)
c. 热焓(5.04%)

(4) Compressor
a. 压缩机(85.44%)
b. 压缩器(13.22%)
c. 压气机(1.15%)

It may be difficult for a seasoned human translator to make a simple decision just on the basis of several alternate renditions. Yet, given the additional relevant usage frequency of each alternate rendition, an initial choice may be made more easily, as most automated processing system would do, by selecting the one with the highest usage frequency.

However, simply relying on statistical distribution may be inadequate for a human translator or an MT system, as can be seen in the case of *aocao* 凹槽 "groove" in Chinese and its alternate renditions in English.

(5) *aocao* 凹槽
a. groove (36.36%)
b. recess (30.82%)
c. grooves (25.78%)
d. notch (3.32%)
e. indentation (1.42%)
f. recessed portion (0.45%)
g. recesses formed (0.37%)
h. trenches (0.34%)
i. concave groove (0.2%)

These examples show that in the actual translation workflow, the usage difference between the non-unique correspondences "groove" and "recess" may be small, and that more than just the identification of alternate translations may be required.

2.2 Value of authentic usage examples

The problem of making an informed decision merely based on the statistics and top choice among alternate renditions cannot be underestimated. Information on actual usage may be needed, as illustrated by two examples below involving authentic alternate renditions.

E.g.	Rend.	%	Context
(6) eutectic point	a. ***gong jingdian* 共晶点**	0.55	如图 1 的相图所示，饱和溶液当冷却时，随着溶液浓度向**共晶点**变化，先沉淀出一种溶质组分。
	b. ***gong rongdian* 共熔点**	0.26	一般来说，许多特性是本发明的去冰组合物所需要的，如低**共熔点**，值接近 7.0 的 pH 和低腐蚀百分数。
(7) *culi* 粗粒	a. **coarse particles**	0.56	However, the method for controlling the **coarse particles** contained in the hard coat layer is not established.
	b. **coarse grained**	0.16	An upper surface of the wafer has sintered thereon a dispersion of **coarse grained** capacitor grade tantalum powder 12.

Table 1. Alternate renditions of *eutectic point* & 粗粒

While E.g. (6) shows that there may not be critical difference in content of the top choice and the second choice, this is not always the case. In E.g. (7), it can be seen that the choice between the two alternate established renditions of *culi* 粗粒 would not be correct if it is determined only by statistics. The correct choice between **"coarse particles"** and "**coarse grained**" has to be determined by the local grammatical context, which is more readily appreciated by the human translator who may not be familiar with the lexical item and the subject, but whose knowledge of grammar would enable him/her to make the proper selection much more easily than an automated system which is statistically bound. Thus, the need is great for the appropriate curation of data to include authentic examples of actual use. This is also true for the case of E.g. (5f) "recessed portion" and E.g. (5g) "recesses formed" for *aocao* 凹槽 in section 2.1.

3. Quantifying and encapsulating multiple renditions

MWEs may pose challenges to human translators in terms of (1) multiple renditions and (2) technicality. We postulate that the extent of cognitive and other efforts required to process sentences or texts with multiple renditions should have a bearing on the difficulty level in translation. The effort and time needed to translate unfamiliar technical terms should also pose challenges in terms of the lexical lookups and decisions to be made. MWEs may provide even more challenges to L2 speakers because they would have less exposure than L1 speakers to very relevant actual language use contexts (Foster et al. 2014, Wray 2002). Similarly, L2 translators may be at a disadvantage when it

comes to phrasal words and MWEs in L2. There are at least two key variables which require elaboration: (1) the extent of the multiple renditions of the term, and (2) the relative importance of the associated terms within the technical area in terms of usage frequency. If the hypothesis stands, then we could have a preliminary measure of translation difficulty by which amount of efforts and relevant linguistic knowledge may be compared.

The measure we postulate, called Multiple-Rendition Index (MRI), quantifies the relative ease of translation between any two given texts on the assumption that there is correlation between the extent of multiple renditions and the difficulty in translation. The MRI could considerate two features: (1) The lexical gravity of the item in terms of its frequency of occurrences and (2) The type/token ratio of the related items within specific domains, and generally. Additional weighting may be provided to reflect the status complexity in lexical registers and other factors. But foremost in the efforts should include a good database.

4. Data cultivation and curation

The alternate renditions and relevant statistics given in this paper are drawn from the Patentlex corpus (http://patentlex.chilin.hk). It is a very large collection of Chinese and English patents which have been found to be comparable, if not parallel in content. It provides the rare golden standards in translation because its cross-lingual terms have been produced by top language professionals and could have legalistic consequence.

Based on a special collection of 10 years of Chinese and English patents, Patentlex has been cultivated specially for NLP applications. It took several years to identify the patents registered under different jurisdictions: in China SIPO (State Intellectual Property Office of China), in Europe WIPO (World Intellectual Property Organization), and in America USPTO (United States Patent and Trademark Office); and it took longer to build up the pairs or sets of comparable patents written in Chinese and English, whose contents are identical or very comparable as determined by NLP means. This has involved culling very large and separated collections of English and Chinese patents (9 billion characters in total) to identify the bilingually comparable patents (Lu et al. 2010). By means of a combination of search efforts, [6] more than 300,000 such Chinese-English patents were identified. We then applied a series of alignment algorithms and found initially 45 million bilingually aligned sentences or sentence fragments, statistically determined to be good candidates of parallel pairs (Lu et al. 2010). These initial sentences were further refined, and provided more than 30 million top quality bilingual sentence pairs. An initial subset of these bilingual sentences was fruitfully used in two pioneering NTCIR Patent MT competition in Tokyo in 2009 and 2010 as a training corpus and then assessment norms (Goto et al. 2011).

It should be noted that the statistically alignment results were basically strings of characters in Chinese and strings of words in English, which may not be all well-formed terms. To obtain linguistically well-formed words or MWEs, further efforts have produced nearly 3 million candidates of bilingual terms so far (Lu et al. 2011a, 2011b, 2010; Tsou et al. 2017, 2019). Currently on-going semi-supervised efforts have yielded nearly one million top quality terms and their multiple renditions used in the analysis reported here.

The production flow for the current corpus of bilingual terms is shown below.[7]

Corpus Cultivation		Corpus Curation		
Stage 1	Stage 2	Stage 3a	Stage 3b	Stage 4
Search 9 billion chars of C&E patents	*Identify* 300,000 comparable C-E patents	*Align*/get 45M C-E parallel sent. pair candidates	*Refine* 30M good C-E parallel sent. pairs	*Filter/get* 3M bilingual MWE candidates

Table 2. Data cultivation and curation of Patentlex

The technical language found in patents is quite representative up-to-date within a specified period. A major difference between the genre of patents and of general texts, is in the vocabulary. Table 3 below provides useful comparison between Patentlex and a Pan-Chinese media report database LIVAC, [https://en.wikipedia.org/wiki/LIVAC_Synchronous_Corpus].

	Doc.-lv.: avg. sent./doc	Sent.-lv.: avg. chars/sent.	Word-lv.: avg. chars/word
CN patent	302.8	54.3	2.12
CN media reports ==> LIVAC	11.5	46.6	1.72

Table 3. Comparisons between pantents & media texts

[6] The primary approach is collocational information, as suggested in Church and Hanks (1990), see also Church (2020).

[7] This collection is bigger than the 7000 preliminary parallel Chinese-English patents reported in Lu and Tsou (2009), as it is much more extensive in size.

Excluding diagrams, an average-size Chinese patent document contains about 300 sentences, which is much longer than the average 11.5 sentences of newspaper texts. More specifically, the average number of Chinese characters per sentence at 54.3 is higher than that of media texts (at 46.6). Also, the average number of characters per word at 2.12 in patents is nearly 25% higher than the 1.72 in Chinese media texts (Tsou & Kwong 2015). It can be readily seen that the compound words and MWEs in patents would outnumber media texts.

By providing both usage frequency and authentic examples, associated with MWEs, Patentlex could assist translators, especially when dealing with multiple renditions. It could also form a basis for MRI which characterizes the translation difficulty for a certain task at hand.

5. Patentlex vs. other corpora

The differences between technical vocabulary and ordinary vocabulary may be also explored. To do this, we compare the technical vocabulary from Patentlex and ordinary vocabulary from LIVAC again in Figure 2.

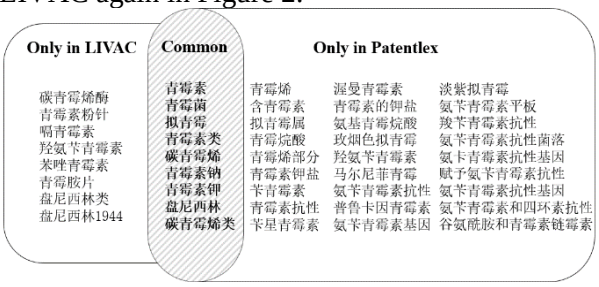

Figure 2. Comparison of the *penicillium* entries from Patentlex and LIVAC

Not surprisingly there are some overlaps. That there LIVAC has some items not found in Patentlex is also not surprising because Patentlex is restricted to a window of 10-year as a retrieval period, in which only new technical developments worthy of protection by law would be reported in the patents. On the other hand, LIVAC reflects topical issues related to the bacteria within their daily life. Notably, a large number of items are found only in Patentlex, which have uncovered just a part of the knowledge base important to our well-being.

Apart from LIVAC, we also compare the Chinese entries containing "青霉" in Sketch Engine and in Patentlex.[8] There are remarkable differences which may result from the data sources of the two corpora. While Patentlex contains 57 terms related to *qingmei* 青霉, Sketch Engine (zhTenTen17) has less (49 items) with 17 overlapping items (see Appendix 1, Table 2 for the examples). However, there is a major difference between the two databases: Patentlex offers also 64 senses in translation related to the 57 terms as well as the distributional statistics of the alternate renditions. Moreover, Patentlex is unique in offering authoritative English translation.

To begin to explore the diachronic development of technical terms, we also compare the Patentlex database with *Modern Chinese vocabulary*, a 1984 Dictionary containing 100,000 entries of Chinese technical terms and ordinary vocabulary. Only 5 terms are found in the 1984 publication, four of which are in common with Patentlex: (a) *qingmei* 青霉 "penicillium", (b) *qingmei-jun* 青霉菌 "penicillium oxalicum", (c) *qingmei-su* 青霉素 "penicillin", (d) *panixilin* 盘尼西林"penicillin", and the fifth is (e) *qingmei-gua* 青霉瓜 "penicillium melon"[9]. This paucity in overlap is not a reflection of extreme developments in the field but of the undeveloped efforts in data cultivation 36 years ago.

6. Cross-lingual MWEs in Patentlex

6.1 Cross-lingual correspondence of MWEs in Chinese and English

The preliminary collection of bilingual technical term candidates exceeds 3 million. While on-going rigorous efforts are made to select the best sets, we can meanwhile report on a preliminary analysis of the lexicons in Chinese and English technical texts on the basis of patents.

Some overall cross-lingual characteristic differences from 100,000 representative items are given below:

	E-to-C	%	C-to-E	%
1 to 1	46843	71.35	74323	85.42
1 to 2	10617	16.17	9302	10.69
1 to 3	3787	5.77	2070	2.38

[8] The source of the Sketch Engine "Chinese Web 2017 (zhTenTen17) Simplified" corpus is mainly the Chinese media and web pages from three Chinese communities, while Patentlex focuses on the patents registered under different jurisdictions.

[9] It has not been possible to trace the source of this term.

1 to 4	1731	2.64	752	0.86
1 to ≥5	2675	4.07	565	0.65
Total	65653	100	87012	100

Table 4: E-to-C & C-to-E multiple renditions

It can be seen from Table 4 that there are noticeable differences among the cross-lingual multiple renditions of the terms. We note that one-to-one translated terms dominate both going from English to Chinese (E-to-C), and from Chinese to English (C-to-E) at 71.35% and 85.42% respectively. Furthermore, 1 to 2 and 1 to 3 alternate translations contribute to the next big group in both directions of translation. It is notable that the percentage of E-to-C multiple renditions (28.65%) almost doubled than that C-to-E (14.58%). Moreover, the correspondence could be as many as 1 to 42 for E-to-C, *and* 1 to 19 for C-to-E cases.

This asymmetry of multiple renditions between the English base and the Chinese base is striking, and invites explanations which may be due to inherent linguistic and lexical differences or due to the direction of translation in the creation of the bilingual documents. If it is the latter, it may be suggested that translation is into a new domain. When there are comparatively inadequate reference materials just as the field of knowledge is developing, there could be many alternate renditions as attempts to create new terms are being made before statistical priorities are established. This is, however, basically a conjuncture which should be evaluated against other more deeply related causes of linguistic difficulty.

As an example, we can compare terms in the field of antibiotics, such as renditions of *penicillium* in Table 5 (see Appendix 1, Table 1 for the examples):

Renditions	E-C	%	C-E	%
1 to 1	17	73.91	26	89.65
1 to 2	3	13.04	3	10.34
1 to 3	3	13.04	0	0

Table 5: Distributions of *penicillium* renditions

For *penicillium*, the gap in asymmetry is larger than the general average in Table 4.

6.2 Network representations of MWEs

The semantically related terms are interrelated and may be represented in a network structure relevant to the mental lexicon. An attempt is made in Figure 3 to show three levels of associated renditions for Chinese *chukou* 出口:

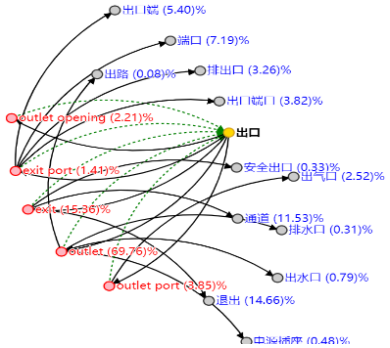

Figure 3: Network representation of *chukou* 出口

In level 2, we have 5 different English renditions, with their usage distributions in the 300,000 patent corpus are given: (a) "outlet" (69.76%), (b) exit (15.36%), (c) outlet port (3.85%), (d) outlet opening (2.21%), and (e) exit port (1.41%).

In level 3, for example, English "outlet" has 4 Chinese renditions (a) *paishuikou* 排水口 (0.31%), (b) *chushuikou* 出水口 (0.79%), (c) *dianyuanchazuo* 電源插座 (0.48%), and the original rendition *chukou* 出口. In addition to "outlet", the other words also have similar multiple renditions. There could be also level 4 and 5. This expanded network of alternate renditions offers a broader view of an important aspect of the lexicological structure of the target terms for the non-causal translators, as well as for the lexicologists and lexicographers.

6.3 Stratification and structure of MWEs

The constructions of MWEs can be quite different in Chinese and in English, as in the *penicillium* case (see Appendix 1, Table 1 for the examples). For Chinese, both semantic adaptations (e.g. No. 3 *qingmeisu* 青霉素 "green-mildew element") and phonetic adaptations (e.g. No. 4 *panixilin* 盘尼西林 "penicillin") are found (and see No. 31 *marnifei-qingmei* 马尔尼菲青霉 "penicillium marneffei" for a hybrid case of phonetic & semantic adaptations). [10] Also the semantic adaptation *qingmei* 青霉 forms the primary base for the majority of terms. The use of Latin and latinate words derived from *penicillium* in English is by far the dominant mode[11]. It can be seen then that the translator from Chinese to English will be at a disadvantage if s/he knows no Latin words

[10] The phonetic mode of adaptation is more common in some Chinese speech communities than others.

[11] See Tsou (2001) and Shen (2001).

associated with bacteria. This is not only found in medicine, but also in law.

The use of Latin or latinate words in English represents an important differentiation between vocabulary in the popular language and that in the High register or learned language. This stratified situation reflects diglossia in general (Ferguson 1959, Tsou 1983). The latinate terms also provide for a common knowledge base for the European languages.

In the case of Chinese, the differentiation in lexical layers to correlate with registers is more subtly manifested through the use of elements from the Classical Chinese language as Table 6 shows:

Class.	***Md.***	***Eng***	***Class.***	***Md.***	***Eng***
fu 腹	*du* 肚	stomach	*lu* 顱	*tou* 头	head
fu 婦	*nv* 女	female	*chi* 齒	*ya* 牙	tooth
kou 口	*zui* 嘴	mouth	*zhi* 脂	*you* 油	fat
zu 足	*jiao* 脚	foot	*yi* 疫	*bing* 病	decease

Table 6: Classical & Modern Chinese morphemes

For the average Chinese speakers, the words from the Classical language are used mainly in professional and learned vocabulary, (e.g. *fuxie* 腹泻 "diarrhoea", *yushi* 浴室 "bath room", *qinshi* 寝室 "bed chamber") and official documents (e.g. *shou* 售 "sell", *gou* 购 "buy") where at least one morpheme belongs to the Classical Chinese language.

However, the use of Classical Chinese elements in a realistically virtual diglossic environment is much more pronounced in the languages of Sinosphere countries, including Japanese, Korean, and Vietnamese, because of their prior incorporation of the Classical Chinese language and the logographic writing system.

The use of Classical Chinese words in these languages serves a function equivalent to Latin and Latinate words in English and European languages and has similarly facilitated intra-regional communication within Sinosphere countries. This deeply rooted tradition also has served to guide the development of new terms.

The diversity in the origin of some Japanese words provides a broader perspective which includes its more recent contact with English. Thus, for describing emotional relationship in Japanese, there can be three basic lexical items: (a) 好き"*suki*", "like" in native Japanese), for a range of casual to deep feelings of positive emotion, (b**)** 愛 ("*ai-suru*", "love" from Sino-Japanese) for a much more serious and intensive feeling, and c) ラブ ("*rabu*", "love" from English for more recent words such as ラブホテル *"love hotel"*, ラブボート *"love boat"*).

It is noteworthy there is two-way flow so that some High register terms in Chinese have come from Japanese through the logographic circle of Sinosphere languages. For example, *hotei (fating* in Chinese*)* 法庭 "court", *minshiu* (*minzhu*) 民主 "democracy", and *sheji (zhengzhi)* 政治 "politics" were first coined in Japan during the Meiji Era before they were introduced into China.[12,13]

The case of "appendectomy" in Japanese and Chinese may provide useful comparison.

(8) *Kyusei-kaifuku-chusui-setsujo*
急性 開腹 蟲垂 切除 (Japanese)
Jixing-mangchangyan-kaidao-shoushu
急性 盲肠炎 开刀 手术 (Chinese)

The Chinese term *mangchang* 盲肠 "blind intestine" refers to the appendage at the end of the digestive track. It first appeared in Japanese as 蟲垂 "hanging worm", a semantic adaptation of the Latin term "vermiform" (worm shape), through the Dutch language, whose speakers along with the Portuguese were the earliest Western visitors to Japan. Lexical stratification is common in languages and serve certain useful social-cognitive functions of differentiation within the society.

Another example is **"subdural hematoma"**, the term referring to "blood clot under the skull",

(9) *Komaku-ka-kesshu* 硬膜下血腫 (Japanese)
Yinnaomo-xia-xuezhong 硬腦膜下血腫 (Chi.)

The English term draws from Latin "sub" and "dura", and from Greek "hematoma", while the Sino-Japanese term refers to "blood swelling under the (hard) skull membrane" [14]. This term would not be easily understood by the man on the street in Japan if it was spoken, but reading

[12] While many words related to modern governance in Chinese have come from Japanese, many words related to cuisine in English have come from French. such as "boeuf" (beef), "mouton" (mutton), and "porc" (pork) as discussed in Section 2.

[13] Similarly, many terms related to Western medicine were first translated in Japan before being adapted in China. This is especially true for matters relating to surgery, such as "appendectomy", which essentially did not exist in China in any significant way. According to Confucian doctrine, the sanctity of body inherited from one's parents should not be violated by unnatural incision. This explains in part why surgery has been a late development in China.

[14] There is simplification in Japanese with the removal of *no* 脳 "brain".

the Kanji characters would improve his comprehension to realize that there is involvement of "brain" and "blood swelling", almost on par with a Chinese man on the street. [15]

6.4 Intra-strata comparison of MWEs

Cross-lingual MWEs may also be compared in terms of headwords. It is noteworthy that while "penicillin" and its derivative terms often function as attributes to other headwords in English, "*qingmei* 青霉" is used more frequently as the headword in Chinese. A good example is "penicillium" in Appendix 1, No. 28 "penicillium citrinum" vs. *juqingmei* 桔青霉.

Another typological difference lies in the number of headwords. A list of 50 common headwords in Chinese MWEs is given in Appendix 2. We note the average number of entries for these top 50 heads is 659, and that the average frequency of occurrence for the 110 top frequency items within the entries at 8100 is quite significant (examples may be seen in Appendix 3). One head may form a number of MWEs in Chinese, but the English headwords involve the use of a large vocabulary of Latin words and display great diversity. The number of headwords in English shall be larger than that in Chinese.[16]

7. Concluding remarks

Among various kinds of MWEs, this paper has singled out bilingual technical terms, purposefully curated from over 300,000 bilingual comparable patents, and has focused on issues related to their non-isomorphic cross-lingual correspondences. We have proposed that the complexity issue may be exasperated by differences in inherent linguistic structure, and that possibly at the inception of terminological development, greater variety and selectivity in the target language may be common when human translation efforts are involved. At the same time, we have pointed out that MWEs in both Chinese and English exemplify passive diglossia with two different lexical layers: the Low language of everyday speech of the population, and the High language known to a much smaller subset of the population. In the case of English, the High register includes Latin or latinate words which are shared by most of the European languages. They also make two important and different contributions: (a) the provision of an easily shared knowledge base in the Western civilization, and (b) differentiation between the status of those who could manage both the Low and High registers in each speech community and those who could not. The same is true of Sinosphere in Asia where some languages have experienced the logographic writing system and the adaptation of some older forms of the Chinese language into their High register vocabulary. This High register layer has contributed to a situation similar to the two-layer system of Europe, with the two different functions. This is also true of speakers of the Chinese language, for whom the contrast between the two registers is less pronounced but no less important. We have also suggested that a Multiple Rendition Index (MRI) measure may be beneficial, and attempted to provide preliminary network representations of MWEs by making use of their multiple renditions and their relative significance within the system, which could have a bearing on the mental lexicon.

Given that technological advancements will always outpace lexicology and lexicography, the integrated study of MWEs through linguistics and natural language processing could go hand in hand to facilitate their management in different applications and our understanding of this important area of language.

Acknowledgements

We wish to thank many individuals who have contributed to the work leading to this publication: Janice Chong, Kenny Mok, Nguyen Thi Hong Quy, Ulrica Nie, Biwei Pan, Belle

[15] There are also instances of portmanteau words in Vietnamese:

(a) Acute appendicitis surgery:
phẫu thuật - viêm - ruột thừa - cấp tính (Vietnamese)
(SinoV.) phẫu thuật 剖術 "operation"; (SinoV.) viêm 炎 "inflammation"; (Viet.) ruột thừa "extra-intestine"; (SinoV.) cấp tính 急性 "acute"

(b) Subdural hematoma:
Tụ - máu - dưới - màng cứng (Vietnamese)
(SinoV.?) Tụ 聚 ? "accumulation"; (Viet.) máu "blood"; (Viet.) dưới "under"; (Viet.) màng cứng "membrane-hard"

Example (a) is more commonly known than (b) and we gather that the latter term is predated by (a). In (a), up to three out of the four constituent words may then root to Sino-Vietnamese words. In the latter case (b), there is only 1 word which might have Sino-Vietnamese origin. These two cases seem to reflect the dwindling use of Sino-Vietnam words in lexical development in Vietnamese, which is different from Japan.

[16] Situations such as this invite comparison between Chinese [Attribute+Head] and English [Head+Attribute] constructions which is beyond the scope of this paper.

Yuan, Skaya Wang, Yuki Wong, and Elaine Zhao.

References

Amelung, Iwo. 2004. Naming Physics. The Strife to Delineate a Field of Modern Science in Late Imperial China. *Mapping Meanings: Translating Western Knowledge into Late Imperial China.* 381-422. Leiden: Brill.

Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. 2009. Statistically driven alignment-based multiword expression identification for technical domains. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 1–8.

Church, K. W. & Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–9.

Church, Kenneth. 2020. Emerging trends: Subwords, seriously? (Keynote Presentation) *The 21st Chinese Lexical Semantics Workshop* (CLSW2020), City University of Hong Kong.

Fellbaum, C. 1998. WordNet: An electronic lexical database. *Language, Speech and Communication*. Cambridge: MIT Press.

Ferguson, Charles A. 1959. Diglossia. *Word* 15(2). 325-340.

Foster, P., Bolibaugh, C. & Kotula, A. 2014. Knowledge of nativelike selections in a L2. *Studies in Second Language Acquisition* 36. 101–132.

Goto, Isao, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin Tsou. 2011. Overview of the patent translation task at the NTCIR-9 workshop. *Proceedings of the NTCIR-9 Workshop*, 559–578.

Goto, Isao, Bin Lu, Ka Po Chow, Eiichiro Sumita, Benjamin Tsou, Masao Utiyama, and Keiji Yasuda. 2013. Database of human evaluation of machine translation systems for patent translation. *Journal of Natural Language Processing 20(1)*: 260-286.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*. Cambridge: MIT Press.

Kwong, Olivia. Y., Benjamin Tsou, and Tom Lai. 2004. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1). 81-99.

Liu, Yuan (ed.). 1984. *Xiandai Hanyu Cibiao* [Modern ChineseVocabulary]. Beijing: China Standard Press.

Lu, Bin and Benjamin K. Tsou. 2009. Towards bilingual term extraction in comparable patents. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation* (PACLIC 23): 755-762.

Lu, Bin, Benjamin K. Tsou, Tao Jiang, Jingbo Zhu and Olivia Y. Kwong. 2011a. Mining parallel knowledge from comparable patents. *Ontology learning and knowledge discovery using the web: Challenges and recent advances*: 247-271.

Lu, Bin, Benjamin Tsou, Jingbo Zhu, Tao Jiang and O.Y. Kwong, 2009. The construction of a Chinese-English patent parallel corpus. *MT Summit XII, 3rd Workshop on Patent Translation*, 17-24.

Lu, Bin, Ka Po Chow, and Benjamin Tsou. 2011b. The cultivation of a trilingual Chinese-English-Japanese parallel corpus from comparable patents. *Proceedings of Machine Translation Summit XIII:* 472–479.

Lu, Bin, Ka Po Chow, and Benjamin Tsou. 2013. Comparable multilingual patents as large-scale parallel corpora. *Building and Using Comparable Corpora (BUCC)* XI: 167-187.

Lu, Bin, Tao Jiang, Kapo Chow, and Benjamin K. Tsou. 2010. Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. *Proceedings of The Workshop on Building and Using Comparable Corpora at LREC-2010*, 42–49.

Luk, Robert, Benjamin Tsou, Tom Lai, O. Y. Kwong, Francis Chik, and Lawrence Cheung. 2003. Bilingual legal document retrieval and management using XML. *Software practice and experience* 33, 41-59.

Ren, Z., Y. Lü, J. Cao, Q. Liu, and Y. Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 47–54.

Sag, I., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: a pain in the neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, 1–15.

Shen, Guowei. 2001. The creation of technical terms in English Chinese dictionaries from the nineteenth century. *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China.* 287-304. Leiden: Brill.

Tsou, Benjamin K. & Olivia Kwong. 2015. LIVAC as a monitoring corpus for tracking trends beyond linguistics. *Linguistic Corpus and Corpus Linguistics in the Chinese Context (Journal of Chinese Linguistics Monograph Series 25)*, 447-471.

Tsou, Benjamin K. 2001. Language Contact and Lexical Innovation. *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*, 35-56.

Tsou, Benjamin K. 2012. Idiomaticity and classical traditions in some East Asian languages. *26th Pacific Asia Conference on Language, Information and Computation*, 39–55.

Tsou, Benjamin K. 2019. From the cultivation of comparable corpora to harvesting from them: A quantitative and qualitative exploration. *Proceedings of the Conference on Building and Using Comparable Corpora (BUCC 2019)*, 29-36.

Tsou, Benjamin K., Derek F. Wong, and Ka Po Chow. 2017. Towards the generation of bilingual Chinese-English multi-word expressions from large scale parallel corpora: An experimental approach. *EUROPHRAS*, 162-168.

Tsou, Benjamin K., Ka Po Chow, Junru Nie, and Yuan Yuan. 2019. Towards a proactive MWE terminological platform for cross-Lingual mediation in the age of big data. *Proceedings of The Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, 21-27.

Tsou, Benjamin K. 1983. Triglossie et realignment sociolinguistique. *Contrastes*. 10-15.

World Intellectual Property Office. 2019. *World Intellectual Property Indicators 2019 -Patents*. Retrieved from https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2019-chapter1.pdf

Wray, A. 2002. *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Appendix 1: MWEs with “penicillium” from different corpora

Table 1: The penicillium words from Patentlex

1	penicillamine	青霉胺
2	penicillic acid	青霉酸
3	penicillin	青霉素
4	penicillin	盘尼西林
5	penicillin acylase	青霉素酰化酶
6	penicillin allergy	青霉素过敏
7	penicillin antibiotic	青霉素抗生素
8	penicillin antibiotic	青霉素类抗生素
9	penicillin antibiotic	青霉素抗生素类
10	penicillin antibiotics	青霉素抗生素
11	penicillin binding protein	青霉素结合蛋白
12	penicillin binding proteins	青霉素结合蛋白
13	penicillin derivative	青霉素衍生物
14	penicillin G	青霉素 G
15	penicillin G	苄青霉素
16	penicillin G sodium	青霉素钠
17	penicillin potassium	青霉素钾
18	penicillin streptomycin solution	青霉素链霉素溶液
19	penicillinase	青霉素酶
20	penicillins	青霉素类
21	Penicillins	青霉素类抗生素
22	penicillins	青霉素类药物
23	penicillium	青霉
24	penicillium	青霉属
25	penicillium	青霉菌属
26	penicillium chrysogenum	产黄青霉
27	penicillium chrysogenum	产黄青霉菌
28	penicillium citrinum	桔青霉
29	penicillium expansum	扩展青霉
30	penicillium italicum	青霉病
31	penicillium marneffei	马尔尼菲青霉
32	penicillium oxalicum	青霉菌

Table 2: MWEs with *qingmei*“青霉” in Sketch Engine (zhTENTEN 2017) and Patentlex

Common	氨苄青霉素 (1818/2643)*, 产黄青霉 (46/64), 桔青霉 (39/13), 拟青霉 (59/102), 拟青霉属(23/67), 羟氨苄青霉素(5/309), 青霉 (1924/995), 青霉胺 (870/408), 青霉病 (205/4), 青霉菌 (1069/11), 青霉属 (115/413), 青霉素 (30886/3753), 青霉素 G (610/352), 青霉素酶 (426/146), 青霉酸 (33/2), 青霉烯 (45/29), 碳青霉烯 (135/103)
Sketch Engine	碳青霉 (939), 碳青霉烯酶 (104), 青霉烯酶 (54), 耐碳青霉 (46), 黄青霉 (40), 青霉索 (34), 青霉菌病 (30), 青霉烷 (21), 蛾拟青霉 (21), 产碳青霉烯酶 (20), 耐碳青霉烯 (16), 青霉烷砜 (15), 绿青霉 (15), 微紫青霉 (12), 蝉拟青霉 (11), 点青霉 (10), 氨苄青霉 (9), 紫青霉 (8), 展青霉 (8), 氨苄青霉霉霉 (SIC) (7), 氧青霉 (6), 青霉类 (6), 苄星青霉霉 (6), 青霉等 (6), 青霉烯酸 (6), 青霉组 (5), 克拉维酸钾/羟氨苄青霉素 (5), 疣孢青霉 (5), 橘青霉 (5), 青霉时 (5), 菌碳青霉 (5)
Patentlex	氨苄青霉素抗性 (393), 青霉素类 (215), 氨苄青霉素抗性基因 (190), 渥曼青霉素 (164), 苄青霉素 (80), 碳青霉烯类 (69), 青霉菌属 (45), 普鲁卡因青霉素 (27),青霉素衍生物 (23), 氨苄青霉素抗性菌落 (20), 扩展青霉 (19), 苄星青霉素 (18),青霉烯部分 (17), 氨苄青霉素平板 (16), 氨苄青霉素和四环素抗性 (15), 青霉素结合蛋白 (15), 产黄青霉菌 (14), 青霉素抗生素 (12), 青霉素链霉素溶液 (12),羧苄青霉素抗性 (12), 青霉素酰化酶 (11), 氨苄青霉素基因 (9), 氨卡青霉素抗性基因 (9), 青霉素钠 (9), 青霉素类抗生素 (7), 氨基青霉烷酸 (6), 青霉烷酸 (6), 赋予氨苄青霉素抗性 (5), 谷氨酰胺和青霉素链霉素 (3), 马尔尼菲青霉 (3), 青霉素过敏 (3), 青霉素抗生素类 (3), 青霉素抗性 (3), 含青霉素 (2),玫烟色拟青霉 (2), 玫烟色拟青霉 (2), 青霉素的钾盐 (2), 青霉素钾 (2),淡紫拟青霉 (1), 青霉素钾盐 (1), 青霉素类药物 (1)

*The numbers refer to the frequencies of occurrence, for common items: (*freq* in Sketch Engine/ *freq* in Patentlex

Appendix 2: 50 Common S/T Headwords from the Patentlex

No.	Headword	Entries
1	器	2022
2	物	1860
3	体	1522
4	(部)件	1275
5	(角)度	1253
6	性	1008
7	量	993
8	剂	980
9	(装)置	962
10	(部)分	931
11	基	772
12	层	738
13	(信)号	730
14	酸	724
15	率	718
16	面	704
17	(系)统	701
18	(分)子	668
19	酯	634
20	(材)料	621
21	(信)息	600
22	(目)的	579
23	(格)式	559
24	线	555
25	化	495
26	(单)元	492
27	点	487
28	(通)道	478
29	数	476
30	构	457
31	力	439
32	(包)括	439
33	素	431
34	(方)法	429
35	(电)流	420
36	(物)质	419
37	(序)列	415
38	(数)据	407
39	(电)路	407
40	胺	396
41	(作)用	395
42	(机)制	391
43	(设)备	388
44	(细)胞	375
45	(类)型	372
46	(区)域	370
47	(反)应	370
48	(组)合	369
49	部	359
50	机	358

No.	Headword	Entries

Appendix3: Examples of MWEs of 10 top frequency Headwords from PatentLex

A. *qi* 器		
English renditions	**Sample entries**	*Freq.*
fully redundant linearly expandable broadcast router	全冗余线性可扩展广播路由器	42
location information domain management server	位置信息域管理服务器	33
fiber bragg grating sensor	光纤布拉格光栅传感器	10
optical recording medium and its corresponding drive	光记录介质及其相应的驱动器	2
complementary metal oxide semiconductor imager	互补金属氧化半导体成像器	1
B. *wu* 物		
acrylate or methacrylate copolymer	丙烯酸酯或甲基丙烯酸酯共聚物	20
aconitrates and citraconates as well as succinate derivatives	乌头酸盐和柠康酸盐以及琥珀酸盐衍生物	20
ethylene alkyl acrylate copolymer	乙烯丙烯酸烷基酯共聚物	2
acrylic emulsions or urethane acrylic copolymer	丙烯酸乳液或氨基甲酸乙酯丙烯酸共聚物	2
ethylene vinyl acetate carbon monoxide terpolymer	乙烯乙酸乙烯酯一氧化碳三元共聚物	2
C. *ti* 体		
diphenylmethane diisocyanate isomers	二苯基甲烷二异氰酸酯异构体	12
cross-linked organopolysiloxane elastomers	交联的有机聚硅氧烷弹性体	10
acrylamide or methacrylamide monomers	丙烯酰胺或甲基丙烯酰胺单体	1
corticotropin-releasing hormone receptor	促肾上腺皮质激素释放激素受体	1
ethylbenzene and all of the xylene isomers	乙基苯和所有的二甲苯异构体	1
D. *jian* 件		
polarization direction rotating elements	偏振方向旋转元件	18
beam shaping optics	光束成形光学器件	18
flip chip semiconductor device	倒装芯片半导体器件	6
optical tool insert	光学加工工具插件	2
conjugated organic semiconductor devices	共轭有机半导体器件	1
E. *du* 度		
low glass transition	低玻璃化转变温度	177
acetylated histone concentration	乙酰化组蛋白浓度	9
bioavailability of metformin	二甲双胍的生物利用度	7
buprenorphine plasma concentrations	丁丙诺啡血浆浓度	2
low crystallinity polymer has a crystallinity	低结晶度聚合物的结晶度	1

F. *xing* 性		
acrylic polymer and a hydrophilic	丙烯酸聚合物和亲水性	3
nitric oxide synthase inhibiting activity	一氧化氮合酶抑制活性	3
dinucleotide repeat polymorphism	二核苷酸重复多态性	2
central dopaminergic neuronal activity	中枢多巴胺能神经元活性	2
acetic acid solution or other acidic	乙酸溶液或其它酸性	1
G. *liang* 量		
comonomer content and molecular weight	共聚单体含量和分子量	3
low density lipoprotein cholesterol levels	低密度脂蛋白胆固醇含量	2
coronary artery blood flow	冠状动脉的血流量	1
average molecular weight of the polyethylene glycols	乙二醇的平均分子量	1
content of vinyl acetate in the copolymer	共聚物中的乙酸乙烯酯的含量	1
H. *ji* 剂		
lipophilic skin moisturizing agent	亲油性皮肤增湿剂	148
phosphite antioxidants	亚磷酸酯抗氧化剂	24
diacylglycerol acyltransferase inhibitors	二酰基甘油酰基转移酶抑制剂	10
ethoxylated alkyl alcohol surfactant	乙氧基化的烷基醇表面活性剂	4
dianionic or alkoxylated dianionic cleaning agent	二阴离子或烷氧基化二阴离子清洗剂	3
I. （*zhuang* 装）*zhi* 置		
portable data storage device	便携式数据存储装置	59
portable inspection data recording device	便携式检验数据记录装置	35
portable radio communication apparatus	便携式无线电通信装置	33
information server memory means	信息服务器存储器装置	13
a portable insulin injection device	便携式胰岛素注射装置	2
J. （*bu* 部）*fen* 分		
erythropoietin portion	促红细胞生成素部分	64
component feed unit control section	元件供送单元控制部分	36
hydrophilic moiety and a hydrophobic moiety	亲水部分和疏水部分	11
donor and corresponding acceptor fluorescent moieties	供体和对应受体荧光部分	1
human monoclonal antibody or a portion thereof	人单克隆抗体或其部分	1