Parallel Corpus in Chinese-English Dictionary Compilation and It's Problems

Yun Hong, Liu Lu
School of Foreign Languages, Sichuan University of Science and Engineering
Zigong, Sichuan, P.R.China 643000

Abstract *- Current methods of compiling bilingual dictionary by corpus have been gaining momentum especially when parallel corpora are developed, breaking many barriers otherwise insurmountable by general corpora. Parallel corpora roll comparative linguistic study, language education and translation into one. Despite their exclusive edges, there are still tricky problems lurking ahead and in sore need of addressing, most of which stem from inadequate study on our own compilation theories and shackles of the past. For instance, due to insufficient use of modern linguistic theories, the compilation process relatively neglects communicative nature of languages. In addition, a wide gap seems to straddle several areas associated with dictionary compilation, making consensus based on our real contexts to equip users with corresponding cultural awareness and international horizons impossible. It's imperative for compilers concerned to expose themselves to more comprehensive prospects around the globe and act in abreast with the call of the times.*

Keywords*: corpus, dictionary compilation, comparative language study*

I. INTRODUCTION

At the turn of 20th century, dictionary compilation has already gone a long way with its origin dating back to 2500 B.C. Concerning the complexity and gravity of dictionary compilation, it's of necessity to point out three integral respects involved: first, the study of meta-lexicography, namely the study of lexicography itself, consisting of collection, interpretation of words, forming sample sentences and sticking to pragmatics, etc. Second, the study of combination of lexicography and pertaining researches of linguistic theories, and third, computer assistance techniques in dictionary compilation.

The most marked change taking place in dictionary compilation is to draw upon corpus. To apply corpus to compile dictionaries is nothing new though in the early 20th century manual collection and relatively random selection pervaded. With computer science gaining momentum and its growingly refined marriage with linguistics in around 1980s, the real sense of corpus-based dictionary compilation has been recognized around the globe. The establishment of COBUILD corpus by Sinclair is considered the milestone in the history of dictionary compilation via computer-accessible corpus. Besides, we cannot afford to neglect the Lancaster University, which as well contributed enormously to this field.

In light of the current situation, national compilers have devised on our own to render corpus-based dictionary compilation convenient and effort-saving. In the thesis, a whole section will be devoted to CpsDict, a self-developed language processing corpus. Nonetheless, there is still a long way to go, and it's essential to discuss problems and conjure up some possible remedies.

In modern sense, corpus used for language research and dictionary compilation is thought to own at least the following three conditions: (1) Representativeness; (2) Purposefulness; and (3) Machine Readability.

The significance of large scale of linguistic material for accurate understanding and describing grammar has been recognized in the late 1950 by Quark. Then he and his team established corpus in its modern sense in London University, i.e. Survey of English Usage, seeking to extensively collect contemporary, social and generic English texts and

speeches. This then translated into the foundation for their grammar treatise with staggeringly convincing descriptions and explanations. Corpus provides linguists with a brand-new means to study language with attention riveted on utilitarian respects and leads to a ground-breaking branch of applied linguistics: corpus linguistics.

II. CATEGORATION OF CORPUS FOR PRACTICAL USES

Monolingual Corpus, the earliest developed and mostly used corpus, collects materials of only one kind of language. It can be further categorized into two kinds: single monolingual corpus that collects original texts of a certain language like BROWN, BNC, and single translational corpus which contains translated texts of a certain language, such TEC.

Bilingual/Multilingual Corpora, clearly by definition, refers to corpus of texts of at least two kinds of languages. There are three forms:

1. Parallel corpora: the bilingual or multilingual corpora consist of both original and translated texts. With regard to parallel corpora, three sub-types also deserve our attention: in uni-directional parallel corpora we can see texts of original language A and translated texts of target language B, and not vice versa. Otherwise, such corpora are called bi-directional parallel corpora in which not only texts of original language A and its translated version of target language B can be seen, but texts of original language B and its translated version of target language A are present. When texts of original language A are translated into language B,C,D..., we label them as multi-directional parallel corpora.

2 .Comparable corpora are comprised of at least two corpora of texts of different languages or varieties of the same language. There exists no translation among their sub-groups, thus making it easy to be insulated from 'translationese'.

3 .Translational corpora features corpora that have translational relationships with each other like TEC(Translational English Corpus).

The benchmarks against which to distinguish the above-mentioned corpora are alignment and translational relations. See the chart below:

Types of Corpora	Translational relations	Alignment
Parallel corpora	✓	✓
Translational corpora	✓	-
Comparable corpora	-	-

It should be pointed out that translation theory research and language comparison research are growingly based on bilingual corpora. As we've mentioned that immune from translationese, comparable corpora are more effective in language comparison in that language A will suffer least negative impact from language B, while parallel corpora excel in figuring out linguistic features and appearance frequency, they are favored in translation study. However, it's advisable to combine both sensibly for translation study and use parallel corpora as a starting point, which propels further analysis about language.

III. CORPUS APPLIED TO SPECIALIZED DICTIONARIES

In a strict sense, terms like corpus-based or corpus-driven dictionary compilation have come under fire from many scholars. Reasons are not hard to find: firstly, corpus is a passive existence for human to explore, incapable of driving or determining contents in the dictionaries, which reminds us of the indispensable roles of compilers and specialized experts in this process. Corpus provides us with language evidence yet also clings to systematic principles. For example, in German, talata refers to Tuesday but its appearance is rare. According to systematic principles, it should be included in conjunction with other words of weekdays despite its low frequency. Second, there is also a dubious attitude towards 'corpus-based' dictionary compilation even though it's dyed-in-the-wool in linguistic and lexicographic field.

Modernized digital dictionaries are compiled through assistance of corpus, suggesting that some parts of this work are done other than corpus. Last but not least, when it comes to 'corpus lexicography' itself, some also argue against it because corpus is just a means for us to avail for lexicographic study. We also use documentaries, questionnaires and dairies for dictionary compilation but never in a time when 'documentary lexicography' or 'dairy lexicography' emerges.

The above-mentioned content has laid a solid foundation for our following discussion about corpus applied to compiling specialized dictionaries, which distinguish themselves from general dictionaries in various ways: involvement of experts, professional explanations of terms, and certain technical contexts against which terms are chosen and used.

Specialized dictionaries were initially compiled for time-pressed or financially disadvantaged job seekers, technicians, practitioners, etc. to find alternatives to keep studying and researching. Some fierce debates flaring up concerns the topic as to who should dominate the compilation process - specialized experts or terminologists? Suggestions are countless like 'dictionary compilation is within the category of lexicographers and linguists.' (Frawley, 1988:196). For us an eclectic attitude is more adoptable: professional experts are places for consultancy while terminologists for professional layout, illustration, and explanations.

It's advisable for us to admit and take account of drawbacks of corpus here applied without dismissing its as whole. From literatures I've recently read, the limitation of corpus in offering interpretations to specialized terms lies in the given 'contexts'. 'Deemed cost', for example, can be explained in two far disparate ways with one in professional accountant dictionary and the other on Google. Laymen like us can hardly tell any differences in meaning from those two explanations but experts can. Those are esoteric words, and the most importantly, contexts. Explanations given in specialized dictionaries stem from experts, who will give a constantly proper context for that specialized term to be used. While in general or online dictionaries, those illustrations are context-dependent, that is when contexts vary so do their specific meanings, which leads to a broader range for the same term to be used.

Corpus in compiling specialized dictionaries has registered its substantial practical value, endowing that process with facts thus adding to authenticity. Lexicographers or terminologists may ascertain the appropriateness of some terms via corpus, but it's not so viable as expected for only experts can figure out whether they are correct or related to some subjects or not with laymen left puzzled and many collocations lost.

IV. PROBLEMS UNDERLYING SLOW PROGRESS IN OUR MISSION

For all staggeringly emerging volumes of dictionaries nowadays, the compilation of Chinese-English dictionaries hasn't witnessed some noticeable advances in theories and technologies. Shortage of lexicographic and lexical theories have rendered dictionary-making out of pace with current global trend. For example, the choice of entries of words and semantic items are largely hinged upon Chinese dictionaries, flaws of which directly dent the quality of Chinese-English dictionaries.

In terms of some national scholars, the use of corpus, or effective use of corpus hasn't been generalized to process of compilation. Problems that are more alarming are as follows:

First, the selection of words items and their interpretation are subjective, and mutual reflection is so severe to the point that negative influences set in. Not only do some unnecessary explanations waste time and impair practicality, but also they are misleading and even too hollow if put into real use.

Second, the choice and order of words senses have no agreed principles. The synthetic induction method employed by a lot of scholars these days is inevitably random, discursive and of blindness.

Other problems include vague or ambiguous compilation purposes, regardless of actively or passively encoded dictionaries; lacking necessary

linguistic information, say, the correct use concrete situations; and improper or gravely repeated illustrations, which have a lot to do with individual judgment.

Without a more micro lens to delve into some loopholes exposed in this process at home and abroad, we still cannot afford to continue discussing details which must be figured out through the dark sides of corpus used in dictionary compilation.

First is whether or nor the corpus itself is typical enough, or does it competent to represent and serve goals. It hinges upon whether its selected materials compliant with objective and scientific sampling principles. According to Biber, 'the chosen should encompass all variables related to study subjects. However, representativeness of corpus are frequently lost to volume (quantities). To shun this situation, there are at least two factors which should be fully considered: the analysis of internal structure. Some national scholars assume that corpus can be seen as a four-dimension model, consisting of time axle, space axle, subject axle and style axle. And size of corpus. The usefulness of corpus shouldn't be mistaken for the bigger the better in that it's more taxing for larger corpus to retrieve and analyze outcomes. Conversely, if smaller corpus is an epitome of quality words or terms and easy to operate, it stands a good chance of disclosing the core essence of overwhelming data, thus outshining bigger one. This is especially true for specialized corpus. In a word, the quality of corpus turns on purposes it serves.

Second is about the scientificity of examples founds to be compatible to target terms. Genuine scientificity will be achieved when four conditions are satisfied: firstly, selected examples can compensate shortcomings of interpretation. It suggests that given staggering disparity between languages, samples should be more context-related and expose learners to use of languages in real life. The second is thought to make learners capable of generalizing what they learn and using them correctly. The third counts its reflection of basic usage and collocations of terms. Last, the fourth is to properly carry certain social and cultural information.

Words of high frequency fall into the third aspect. Though dictionary compilation more and more pragmatic-oriented, no doubt a good sign, compilers and many learners have blurred the borderline between words of most use and words of most times. Such a false idea is not rare and has to some extent opened Pandora's Box in that people are inundated in words unwanted and misleading. For example, the results of corpus indicate that the word 'bid' has surfaces 226 times. But it cannot stand closer scrutiny which collaborates that sources of those words are all linked to some 'auction' articles, and latter it turns out that in general conversations, not once of 'bid' emerges. Considering this, some propose 'degree of use' or 'degree of general use' to depose words of high frequency.

V. THE MAKING AND APPLICATION OF PARALLEL CORPUS

It can be said that the establishment of parallel corpus is the new beacon for comparative language study. Far beyond normal functions of general corpora, parallel ones exert their own unique influences on dictionary compilation, like selection of corresponding words, samples of dictionaries, collocations and culturally-bound terms. From my own point of view, it is 'comparison' between two or more languages exclusive to parallel corpus set it apart and make it realize so many purposes in a more satisfying way. Still, the pure 'equivalent words' should be carefully tackled in that there is a gap between languages in cultural, historical or racial aspects. Blindly seeking equivalence through comparison in parallel corpus will reduce us to be slaves to machines without thinking in a fluid context.

To make parallel corpus, there are three aspects deserve attention: labeling, alignment and application.

(1). Labeling:

The aim of establishing bi/multilingual parallel corpus is to collect and store mutually translational information in practical linguistic communication, and to offer language study and other relevant

research the methods and channels for retrieving that information. What is the key to providing such information, or the medium is raw text corpus. It plays a big role in registering frequency of words and their collocation relationships while it is not strong enough in retrieving information from a more complicated context. Therefore, when building up corpus, the essential step after collecting raw texts is to draw upon some signal system to label information that are predicted to be useful. Obviously, it is no simple task in that it touches upon many other academic fields, such as analysis of language structure, language model and the design of implementation plan. The labeling of corpus can be conducted in terms of different levels. Suffice to say that corpus labeling has become an important branch of corpus research.

(2). Alignment:

One purpose we cannot afford to ignore is to garner mutually translational information. Usually, the relevance degree of semantics in expressions of two or more languages and its local context should be considered when we do translation and language comparison research. Thus we need to first discern the corresponding context of each pair of words translated in different languages. The most common way of alignment is based on sentences.

(3). Application:

With corpus research getting deeper, parallel corpus is gaining growingly extensive application. These days, the major fronts in which corpus finds its way are: corpus linguistics, retrieval of parallel corpus, compilation of bilingual dictionary and Language Engineering, such as machine translation. Specifically speaking, in parallel corpus stores a variety of translational examples in authentic communication, ranging from written to spoken language and ending up being basis for comparable linguistics and translational theory. Yet it's not hard to anticipate that parallel corpus will open the floodgate of comparable linguistics in the foreseeable future.

In the procedure of using parallel corpus for practical ends, researchers are constantly bombarded by problems, among which the noticeable one counts development or attainment of processing program and apps of parallel corpus. Another one is interdisciplinary cooperation because linguistics theories are needed and the same is true for media and expertise in probability statistics. Moreover, it's likely that one individual will not accomplish his task even if he is a master in his profession unless he collaborates with other organizations or agencies devoted to the relevant issues.

Above we have briefly discussed problems exposed in bilingual dictionary compilation, especially Chinese-English dictionaries. But the point now is how parallel corpus can do a better job in fixing problems. In terms of limited zone, the core part will cite translation of cultural items to elaborate on this issue.

In many Chinese-English dictionaries, the translation of '端午节' is 'Dragon Boat Festival'. However, when you read the following text, you may feel perplexed: Since then, on the fifth day of the fifth lunar month every year, each household in China would make glutinous rice dumplings and eat them to commemorate the great poet Qu Yuan. This is a traditional Chinese festival known as 'Dragon Boat Festival'. It's hard for a non-Chinese to know dragon boat if the context given shows no clues about it. Some scholars point out that such cultural items steeped in connotation should be translated literally as 'the Double Fifth Festival'. Failed attempt to display connotation may result from translation that is too specific to a certain event to make sense in general, and word-for-word translation from Chinese.

This is where parallel corpus is called out. With its exclusive edges in comparative linguistic study where translation is the big head, it's able to figure out the most useful expressions needed. Four kinds of translations are summarized as: word-for-word translation, literal translation which fully shows the flavor of source language, free translation which is objective and expressive, and functional equivalence translation where target language is given priority to make locals understand. Based on the compared statistics, parallel corpus will accurately explore a

series of interpretations of different kind which then are tailored for users rather than distilling them with overwhelming but unnecessary materials.

VI. CONCLUSION

Corpus has revolutionized how bilingual dictionaries, especially in our country the Chinese-English dictionaries, and also streamlines multiple aspects involved in that process. Still, it doesn't mean that we should view it as we want in that the principle in dictionary-making by virtue of corpus is that it should be based on corpus rather than being tethered by it. No matter how scientific, reasonable the design and building of corpus is, to blindly depend on corpus will yield eccentric outcomes.

Some research methods based on linguistic competence should be complementary rather than mutually resistant to manners in which corpus is used for focusing on language use.

The core of bilingual dictionary compilation is the careful study words meaning. But due to the particularity of the study of words meaning, it's very essential to allow personal judgment and proper intervention.

As of today, computer science has made inroad in various aspects of our lives. Neglecting its edges, we're doomed to miss a good chance of stepping into a new phase of linguistics, which then extends to lexicography and study of dictionaries. Application of parallel corpus to dictionary compilation has showed our growing awareness towards language use, and more than that, those dictionaries are expected to reveal social, cultural and historical significance to knit the whole world by relating human minds.

This work is supported by Sichuan Federation of Social Science Associations（SC18WY024）.

REFERENCES

[1] Biber D. Representativeness in Corpus Design[J]. Literary and Linguistic Computing, 1993,8(4): 243-257

[2] Ding Dongmei. The Influence of English-Chinese Parallel Corpus on Bilingual Dictionary Compilation [J]. Career Circle. 2007(18):86-87

[3] Frawley W. New Forms of Specialized Dictionaries. International Journal of Lexicography, 1988(3): 189-213

[4] Kang Shiyong，Wang Xinglong，Xie Xiaoyan. Preliminary Investigation and Research on Computer-aided Dictionary Compilation System in China [J]. Journal of Lexicography，2012（3）：6-14

[5] Li Dejun. Some Thoughts on the Application of Corpus in the Compilation of Bilingual Dictionaries[J]. Journal of Lexicography，2006（2）：104-109

[6] Liang Shiting. A Review of Lexicography Studies in China within the Recent Decade [J]. The Knowledge of English, 2015(4):76-79

[7] Wang Kefei. Development and Application of Bilingual Corresponding Corpus[M]. Beijing: Foreign Language Teaching and Research Press，2004