ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 4, No. 1, pp. 46-51, January 2014

doi:10.4304/tpls.4.1.46-51

A Corpus-based Machine Translation Method of Term Extraction in LSP Texts

Wei Huangfu
School of Foreign Languages, North China Electric Power University, Beijing, China

Yushan Zhao
School of Foreign Languages, North China Electric Power University, Beijing, China

***Abstract*—To tackle the problems of term extraction in language specific field, this paper proposes a method of coordinating use of corpus and machine translation system in extracting terms in LSP text. A comparable corpus built for this research contains 167 English texts and 229 Chinese texts with around 600,000 English tokens and 900,000 Chinese characters. The corpus is annotated with mega-information and tagged with POS for further use. To get the key word list from the corpus, BFSU PowerConc software is used with the referential corpora of Crown and CLOB for English and TORCH and LCMC for Chinese. A VB program is written to generate the multi-word units, and then GOOGLE translators' toolkit is used to get translation pairs and SDL trados fuzzy match function is applied to extract lists of multi-word terms and their translations. The results show this method has 70% of translated term pairs scoring 2.0 in a 0~3 grading scale with a 0.5 interval by human graders. The methods can be applied to extract translation term pairs for computer-aided translation of language for specific purpose texts. Also, the by-product comparable corpus, combined with N-gram multiword unit lists, can be used in facilitating trainee translators in translation. The findings underline the significance of combing the use of machine translation method with corpora techniques, and also foresee the necessity of comparable corpora building and sharing and Conc-gram extracting in this field.**

***Index Terms*—term extraction, comparable corpus, GOOGLE machine translation, fuzzy match, language for specific purpose**

I. Introduction

Automatic extraction of terms, especially if they are multiword expressions (MWEs), has come to a growing thorny problem for the natural language community and corpus linguistics. Indeed, although numerous knowledge-based symbolic approaches and statistically driven algorithms have been proposed, efficient extraction method still remains an unsolved issue (Li, 2010). This paper is to examine the possibility of refining term extraction methods by combining machine translation and corpora, and try to make full discovery of how to coordinate between them, and exploit the way of using these tools, which are with a highly complementary functions, to fulfill the proximal recall of bilingual term translation pairs.

II. Discussing the Corpus-based Methods

Corpus linguistics has come to be beyond the sense of methodology for conducting language research, but also a new research domain as a theoretical approach to the study of language (McEnery, 2007). The fast growing volume of corpus and the increasing sophisticated analytic techniques have brought about the fundamental changes to language research. As to the applications, there could be unlimited uses of corpus in all fields of linguistic research, natural language processing, and etc. For this research in particular, right type of corpora have to be designed and used for extracting terminologies in translational studies.

A. Designing the LSP Corpus

There are different types of corpora, considering the language, contents, structure, times, tags and annotations. The common types include general corpora, historical corpora, specialized corpora, learner corpora, speech corpora, multimedia corpora and parallel and comparable corpora, etc. For designing this research, it is better to distinguish between comparable corpora and parallel corpora. As described in Baker (2004) and McEnery (2003), a comparable corpus includes similar texts in more than one language, between which there are similar criteria of text selection though that similarity can vary greatly in researchers' own regards, and a parallel corpus contains texts that are produced simultaneously in different languages, or in another word there are source texts and their translations. Comparable corpora are most suitable for this research, for there are no negative influence of translators in parallel corpora and both English and Chinese texts will be of genuine language uses.

In this research, the size of the needed corpus is around 600,000 English tokens and 900,000 Chinese characters. Though, some web retrieval tools or web spider/crawler programs can be used to download materials from the Internet automatically and quickly, and save a lot of manpower, the texts gathered in this way are of low expected quality and need a large amount of filtering and selecting work. So the researcher has chosen to manually search and download the materials for building the corpus, which is slow but of high cost-efficiency. In the corpus building process, Notepad++ and Editpad Pro are used to clean-up the texts because they are supported with powerful regular expressions and batch processing. Finally, Stanford Tagger is used in segmenting, POS tagging, and de-tokenization.

To keep the external validity of the texts in corpora, the LSP (language for specific purpose) texts are limited only to contracts, and GOOGLE and BAIDU search engine are used to retrieve texts on the Internet by inputting search terms with obvious field domain markers and time restrains in contracts' texts, which are then downloaded for further processing and included in the corpus. To balance the English and Chinese texts in the comparable corpus, English and Chinese texts of the same topic will be selected in a 2:3 proportion, i.e. in general there will be 167 English texts and 229 Chinese texts. The frequency distribution of English text tokens and Chinese text characters are as in Fig. 1.:

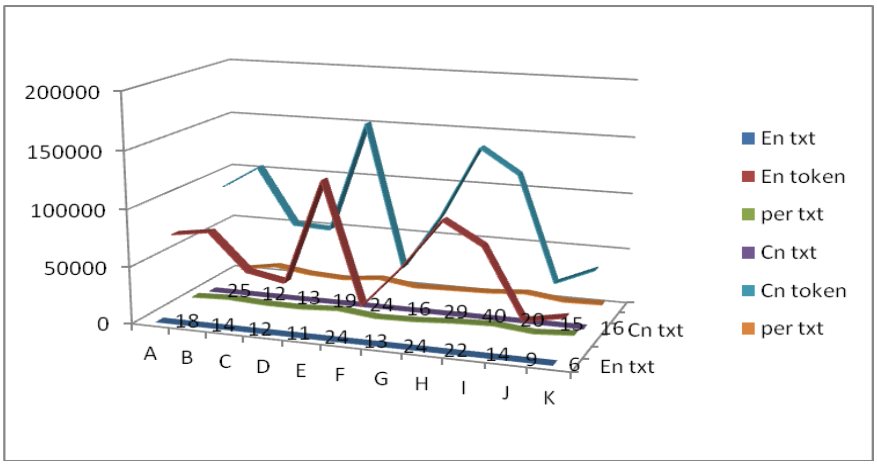


Fig. 1.:A~K represent categories of texts based on contract topics, A is information science and technology cooperation agreement; B is construction engineering projects; C is management and business; D is after-sale service; E is finance; F is confidentiality agreement; G is leasing agreement; H is authorization document; I is purchase agreement; J is employment agreement; K is patient fact sheet

B. Key Words Extraction

Corpora with parallel bilingual texts or multi-lingual texts are usually not large-scale general corpora; rather they are about specialized language uses. Related researches could be the comparison of text features of specific type of genres, or extraction of the translation pairs, etc. Among them is key word extraction, an essential part of the techniques used in text mining.

One text feature in specialized language uses is keyword list. Key words are those that show "aboutness" of texts (Scott and Tribble, 2007). Thus, key words are not necessarily those with high frequency, but those in significantly different use frequency, either particularly higher or lower in use. Key words undoubtedly show the language features of certain domain.

A popular algorithm for key word indexing is the TF-IDF method, i.e. term frequency–inverse document frequency, a numerical statistic which reflects how important a word is to a document in a collection or corpus. As reported in Matsuo (2004), the TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to remove those unwanted and make efficient recall of authentic key words for the fact that some words are generally more common than others. Another method is N-GRAM, i.e. a contiguous sequence of n items from a given sequence of text from a corpus (Ohsawa, 1998). These two methods both use a corpus, and they differ in that TF-IDF method remove stems and stop-words first, while N-GRAM calculate the weight of strings or words of certain length. They use log-likelihood or X^2 as the measurements to decide the degree of bias of the co-occurrence distribution in one corpus and another reference corpus.

Now corpus tools can be used to find the key words in texts by automatically calculating the log-likelihood or X^2 if the observed and referential corpora are available for use. These tools include AntConc, BFSU PowerConc, etc. In this research, BFSU PowerConc is used to extract word lists from both the observed corpus and referential corpus, namely the words and their frequencies, then find out the key words with the log-likelihood (LL) or X^2 measures (Xu, 2012). The larger the LL or X^2 value, the significant use of the words in the observed corpus compared with the referential corpus. In this research, some general corpus, such as the now-popular Crown and CLOB (The 2009 Brown family corpora) for English, and TORCH Corpus and LCMC for Chinese are used instead of a corpus of language use for special purpose, because the contracts vary greatly in language use, i.e. the word uses are not limited to a specific language genre. Their

total size is about 8,000,000 English tokens or Chinese characters, which is, more or less, five times the size of the comparable corpora used for this research.

C. Multi-word Unit Extraction

There are great amounts of literature on the study of multi-word units (MWUs) and so its operating definitions also vary greatly. According to Prentice (2011), scholars have put forward more than 61 terms that are synonyms to multi-word unit. One clarification should be made that not all co-occurring words are formulaic and chunks can also have slots and often are not continuous sequence in natural languages. Compared with collocation, the operating definitions of MWUs are not precise and thus most researchers suggest a combined method of manual and computer-aided coordination. Wordsmith Tools (available from http://www.lexically.net/) have a cluster function program that can be used to generate multi-word units, but it cannot be adapted to extract MWUs in an assigned place within a sentence in a target corpus. Additionally, statistical methods cannot be usable because they will over-generate acceptable strings when comparing co-occurrence of words in prefabricated strings with their separate occurrence in a corpus. So to get rid of the inefficiency, these using tools should be held with caution due to the heavy manual correction work afterwards. A considerable manual work may be necessary to eliminate those unwanted results. Therefore, a novel method has to be researched.

In this study, the starter words have already been generated in the key word extraction process described in the previous part. So, a program is written with the Visual Basic language and the Natural Language Toolkit (NLT), which can be easily accessed via the Internet from http://ishare.iask.sina.com.cn/download/explain.php?fileid=24767255 and http://code.google.com/. One should also not rely too much on computer tools and should be cautious of using statistical measurement values, such as log-likelihood and X^2 and mutual information and T-scores. The MWUs may not the same if using different measurements, and so it may be advisable to use both measurements to find those that frequently appear in each list.

III. Discussing the Machine Translation Methods

Makoto Nagao(1984) first proposed the example-based machine translation (EBMT) method, proposing that translation can be better done with machine by segmenting sentences into translation units, such as clauses and phrases, and then these segments are restructured and translated. Based on this theory, machine translation is generated from the examples in bilingual corpus. But, the EMBT method needs a considerably large reference corpus with bilingual sentence examples, which need huge human and financial resources to build. If the coverage and size of the corpus is limited, computers cannot find perfect matches for the translation. It is for these reasons that the EMBT method may well be used as a plug-in method to improve efficiency and quality of human transition rather than replace it(Dietzel, 2009).

GOOGLE translators' toolkit is such a powerful machine translation tool in that this system has stored a vast amount of bilingual or multilingual translation pairs, which can also be taken as a good way to find equivalent pairs for linguistic constructions in comparable corpora. Moreover, many CAT tools, such as SDL trados, dj à Vu, wordfast, etc., are equipped with fuzzy matching function, which can be made the best use of if there is a translation memory available for comparing the similarities rate between the target and source translation units. Thus, MWUs generated from the comparable corpora will be first processed by GOOGLE machine translation to get translation pairs for both English and Chinese MWUs, and then be made into a translation memory, which will be used by a CAT tool to retrieve the term pairs with the closest similarities from TM. The results will be evaluated by graders to assess the accuracy of recall.

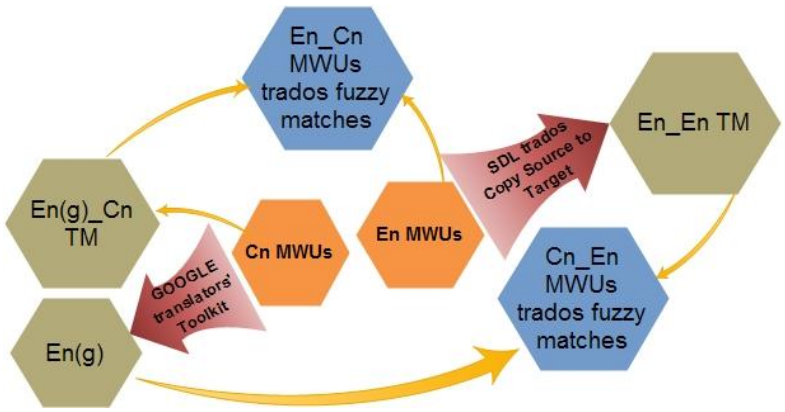


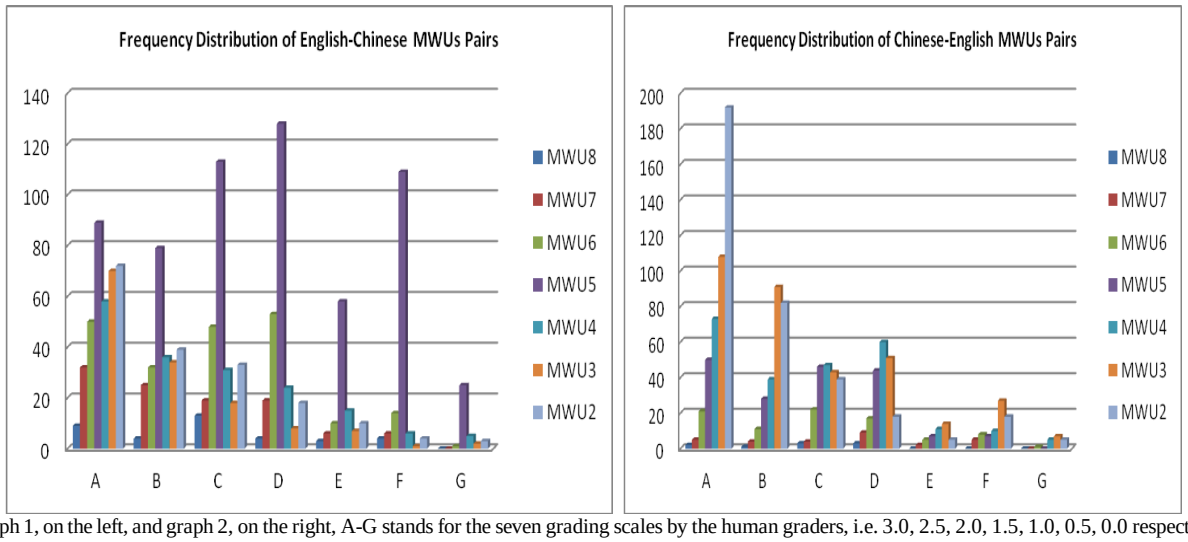
Chart 1: En and Cn represent the English and Chinese languages respectively; En(g) is the Google translation results of Chinese MWUs; E(g)_Cn TM and En_En TM are two MWUs' translation memory

This flow chart demonstrates the procedure of how to generate English-Chinese MWUs' translation pairs and Chinese-English MWUs' translation pairs. The Chinese MWUs are first translated into English via Google translators'

toolkit, and then made into a TM, which will be used to generate fuzzy matches for English MWUs with thresholds set as a range with different values from 75%, 80%, 85%, 90% and 95%. To get the translation pairs of English MWUs, a English-English TM is first built by copying source into target in SDL trados, and then this TM, containing nothing but the English MWUs, is applied to the process of finding fuzzy matches with a threshold set as a range with different values from 75%, 80%, 85%, 90% and 95% for En(g), which is the Google translated result of Chinese MWUs. After both process, two lists of MWUs translation pairs have been achieved.

IV. Results and Discussions

The results from the MWU translation pair extraction process will be put into Excel for manual evaluation and selection. A seven grading scale is used by the human graders, i.e. 0-3 points at a 0.5 interval. The Pearson correlation co-efficiency for the graders are assessed after the grading process, which reaches as high as 0.78 though with a slight difference in assessing English-Chinese and Chinese English MWU translation pairs. The results are presented in two diagrams as follows:



In graph 1, on the left, and graph 2, on the right, A-G stands for the seven grading scales by the human graders, i.e. 3.0, 2.5, 2.0, 1.5, 1.0, 0.5, 0.0 respectively. MWU 1-8 are the multi-word unit translation pairs.

According to the graph 1, i.e. the frequency distribution of English-Chinese MWUs pairs, MWU 5 makes up the largest portion in every grading scale. This makes sense in that five-word MWUs constitute almost 45% of the total MWUs extracted from the corpora. Also this implies that necessary caution should be taken to select good matches from the five-word MWUs translation pairs. It is good to see that MWU 4, MWU 3, MWU 2 produce the most desirable results. These matches are high in quality and also great in number of recall.

In graph 2, i.e. the frequency distribution of Chinese-English MWUs pairs, MWU 4 has the high recall rates but the corresponding accuracy is also low. This can be explained by the frequent use of four-character words in the Chinese language. Strikingly new is the large recall of MWU 3 and MWU 2 with also high accuracy. This implies that for Chinese to English translation pair extraction, two and three character words can produce the most desirable results. Those exceeding four characters, such as those MWUs with 5-8 words are not useful in terms of cost-efficiency.

Moreover, among the total number of 1326 English-Chinese MWUs translation pairs and 1154 Chinese-English translation pairs, those produced in 95-100% fuzzy matches and scoring more than 2 in human assessment both make up around 40% of the total recall, and those in 75-94% fuzzy matches and scoring more than 2 in human assessment both make up also around 30% of the total recall. Though the dispersion of the distribution tilts sharply to the 95%-100% and 75%-80% fuzzy matches, i.e. those in between are both low in frequency of occurrences and in quality. This further proves the methods used in this research are applicable and worthwhile. But, focus need to be on improving the accurate recall of 2, 3, 4, and 5 words or characters MWUs.

Take a close look at the results, and one will find that there are several matches in Chinese translations for English terms, thus should not limit their choices to only one at all. Though the extracted matches are not 100% terms, they are useful when added to translation memory and providing translators with suggestions on their translations. Also, those within 75-94% fuzzy matches are acceptable translations since human graders' evaluations provide further evidence of their usefulness.

However, some discussions should be made on the incorrect matches as is indicated in human graders' assessment. The mismatches are largely due to the shortcomings of Google machine translation, since correct translation pairs still cannot be accessed in its vast example data base due to the following factors. Firstly, two-word units are too short to

provide clear context or contain functional words. These can be avoided to a certain extent if a stop list can be premade to eliminate some results, but the precision and recall are on a contrary relation and a balance is not possible in this sense. Secondly, MWUs of passive or negative structure can result in many mismatches in Google translation, since the machine translation system will omit the words of negation and take passive voice for active voice, which will both cause wrong translations. Thirdly, some Chinese prefabricated chunks are hard to process for Google machine translation system. Comparing the following two pairs, "non-agreed to by Party B of, without the consent of Party A" and "Party B's prior written consent, without the consent of Party A written", the first one is wrong and second one is correct, but both are provided for two MWUs with only one Chinese character difference. Fourthly, there are clear intra-language differences for collocates with high mutual information with the frequently used terms in Chinese and English contracts text. By searching the comparable corpus, one will find there seam not significant overlap of collocate uses for terms about "termination, survival, court, damage and harm, invalid and confidential", etc., and the sentence structure to express these concept in contexts also vary greatly in sentence length and word variety. Deep level structures instead of the word level segments intra-language differences can account for these difficulties in having correct Google machine translation results.

In all, this research has combined machine translation technique with fuzzy match function of computer-aided translation memory system, in which similarity calculation and automatic translation can be both accomplished easily. Thus, this research is to apply simple theory and easily-accessible tools to make full use of comparable corpus in translation term extraction, through which the disadvantages of the now-popular method of using parallel corpus in term extraction can be avoided.

V. Applications and Suggestions for Future Research

The methods to extract MWUs translation pairs used in this research are useful in many aspects. This is another contribution to computer-aided translation (CAT) in that it can provide multi-word units alignment, which is larger than some terms and smaller than sentences. According to Koehn (2003), these MWUs translation pairs are better reference sources in CAT in terms of storage and retrieval efficiency. Among the extracted MWUs translation pairs in this research, there are terms, phrases, and large number of prefabricated chunks. These pairs can be directly used in computer-aided translation if properly built into a term memory in TMX or tab-separated txt format. Moreover, similar to those described in Castillo (2010), these methods of combing MT with corpus, if properly handled, can generate useable results for natural language processing specialists as training materials for improving machine translation system and then apply the machine translation methods to expand parallel corpora.

The proposal of using corpus in translation has long been made as early as in Baker (1993, 1995, 1996) and Laviosa (1998). But as indicated in the research by Garcia (2010), machine translations, such as Google translators' toolkit, trainee translators indeed benefit from the recommended translations. The present research is of great significance in that it has investigated the possibility of how to include such machine translation and make the best of the modern computer-aided translation tools and natural language processing methods. By the way, the subsequent corpus built with these methods for the present research is efficient to be used in preparatory training classes for translators working for a language-specific domain. Trainee translators can use machine translation to seed the empty segments in their computer-aided translation tools and can frequently search with key words for asserting their uses in sentence contexts in the corpus to achieve better translation results.

For future research, an approach should be found to make the comparable corpora of language for special purposes (LSP) available for research uses. It is time-consuming and financially unaffordable to build comparable corpora for every LSP area. For this research, the corpus cannot be made public with immediate access because of copy right protection, which remains a thorny issue to be settled. Another technique problem still poses considerable difficulties to expand the research from N-gram to Conc-gram dimension, which is considered a trendy issue in multi-word unit extraction (Greaves, 2009). Google machine translation system used in this research also produce wrong results for many MWUs due to that its accuracy for translating Chinese are heavily influenced by the correct segmentation of Chinese words, which is a necessary step in this research. Researches have confirmed that it is still difficult to attain 95% accurate segmentation of Chinese (Huang, 2007). And also the currently available corpus tools are not compatible to process Chinese Conc-gram, which makes the expansion of the present study difficult. These factors make it hard for ordinary translators to extract translation term pairs, but still keep the door open to those who know translation well and also are good at programming and other advanced computer uses. So, in the future, another work could be on improving the accurate segmentation of Chinese words or using alternate way of using machine translation system to use Conc-grams in generating translation pairs for multi-word units.

Acknowledgement

This research is financially supported by the Fundamental Research Funds for the Central Universities in China (Grant NO. 13MS47), and by the Young Talents Program of Higher Education in Beijing.

REFERENCES

[1] Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In Sinclair, J. M., Baker, M., Francis, G., & Bonelli, E. T., *Text and Technology: in Honour of John Sinclair*. Amsterdam: John Benjamins Publishing Company, 233-250.

[2] Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2), 223-243.

[3] Baker, M. (1996). Corpus-based Translation Studies: The Challenges that Lie Ahead. Amsterdam: John Benjamins' Translation Library, 18, 175-186.

[4] Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9(2), 167-193.

[5] Castillo, J. J. (2010). Using machine translation systems to expand a corpus in textual entailment. In the *Proceedings of the 7th International Conference on Advances in Natural Language Processing*. New York: Springer US, 97-102.

[6] Dietzel, S. (2009). Example-based Machine Translation. Berlin: Springer Verlag.

[7] Garcia, I. (2010). Is machine translation ready yet? *Target*, 22(1), 7-21.

[8] Huang, C., & Zhao, H. (2007). Chinese word segmentation: A decade review. *Journal of Chinese Information Processing*, 21(3), 8-20.

[9] Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *the Proceedings of the 2003 Conference of the North American Chapter on Human Language Technology.* Massachusetts: Association for Computational Linguistics, 48-54.

[10] Laviosa, S. (1998). The corpus-based approach: A new paradigm in translation studies. *Meta: Translators' Journal*, 43(4): 474-479.

[11] Li, B. & E. Gaussier.(2010).Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *the Proceedings of the 23rd International Conference on Computational Linguistics*. Massachusetts: Association for Computational Linguistics, 644-652.

[12] Makoto Nagao. (1984).A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji, *Artificial and Human Intelligence*. Amsterdam: Elsevier B.V., 173-180.

[13] Matsuo, Y., & Ishizuka, M.(2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1), 157-169.

[14] McEnery, T. & Andrew Wilson. (2003). Corpus Linguistics. UK: Cambridge University Press.

[15] McEnery, T. and X. Zhonghua. (2007).Parallel and comparable corpora. *Corpus-Based Perspectives in Linguistics*, 6, 131.

[16] Ohsawa, Y., Benson, N. E., & Yachida, M. (1998). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of the Advances in Digital Libraries Conference*. New York: IEEE, 12-18.

[17] Prentice, M.(2011). A method for extracting formulaic sequences from a student corpus. *Kanagawa University Language Research*, 34, 35-52.

[18] Scott, M., & Tribble, C. (2007).Textual patterns: key words and corpus analysis in language education. *TESL-EJ*, 11, 2.

[19] Greaves, C., & Antiquariat, J. B. (2009). ConcGram 1.0: A Phraseological Search Engine. Amsterdam: John Benjamins Publishing Company.

[20] Xu, Jiajin, Maocheng Liang & Yunlong Jia. (2012). BFSU PowerConc 1.0. Beijing: National Research Centre for Foreign Language Education, Beijing Foreign Studies University.

Wei Huangfu, the corresponding author, is currently a senior lecturer in North China Electric Power University. He is also the co-ordinator of the College English Teaching Division of the School of Foreign Languages. He has been honored as one of the teacher candidates of the Young Talents Program of Higher Education in Beijing, and has authored or co-authored many academic articles and several books on language education and translation and now is researching in the field of corpus and translation.

Yushan Zhao, the co-author, is currently a professor in North China Electric Power University. She is also the executive Dean of the School of Foreign Languages. Her research interests are in translation and MTI education.