ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 1, No. 3, pp. 287-291, March 2011
© 2011 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/tpls.1.3.287-291

Corpus-driven Learning in Collegiate Translation Course

Qing Wang
Shandong Jiaotong University, Jinan, China
Email: wangqing6906@126.com

***Abstract*—Translation teaching methodology in China needs innovation and corpus has found an increasingly important application in translation studies because its large amounts of stored data of naturally occurring language can enhance learner awareness about the language. The present paper aims at investigating the role of corpus-driven learning in student translator training at the university level in China.**

***Index Terms*—corpus, corpus-driven learning, student translator training**

I. Introduction

Student translator training is troublesome for a number of reasons. Firstly, translation is usually considered difficult for ESL/EFL learners because it involves an important factor: the transfer of mother tongue, which shapes learners' inter-language. Oldin (1989) asserts that transfer has been documented to occur at all the levels of linguistic analysis such as phonology, syntax, lexis, and grammar. Towel and Hawkins enumerate five observable phenomena about second language acquisition of which transfer of L1 patterns into L2 is of prime significance: "Transfer seems to affect all linguistic levels: pronunciation, syntax, morphology, lexicon and discourse." (Towel and Hawkins, 1994, p.7) Translation into a foreign language involves all these aspects of transfer, demanding a comprehensive capacity on the part of the student translator.

Secondly, in the context of Chinese collegiate curriculum, the credit hours of the translation course are far from adequate. In Shandong Jiaotong University, a non-key university in China, the English majors are allocated only 72 credit hours for both English-Chinese and Chinese-English translation, theory and practice all inclusive, lasting only 2 semesters. In Shandong University, one of the key universities in China, things are in a better condition, with the credit hours of translation totaling 144, lasting 4 semesters. In both universities translation, however, is given fewer hours than comprehensive English and advanced English. Considering the complex factors involved in the translation process, the long history of development of translation theories and the vastly diversified discourses of translation practices, such an allocation of classroom hours is all too inadequate.

Thirdly, teacher of translation in China are almost exclusively those whose native language is Chinese and whose English is mostly acquired in the Chinese context, from primary school up to collegiate level. When native speakers of English find it difficult to explain why one word is preferable than the other in certain circumstances and have to resort to language instinct, Chinese teachers teaching C-E translation are faced with an even greater challenge since their feel of the foreign language is far less reliable than native speakers.

With these reasons confronting the student translator training in China, the quality of the trainees falls short of satisfaction. One possible way out is to find a teaching methodology that can help teachers to teach with more confidence and students to translate with more efficiency and accuracy. The recent corpus-driven learning is a possible solution, as it emphasizes the constructive learning process with the support of large quantity of naturally occurring language materials. The present paper aims at investigating the role of corpus-driven learning in student translator training at the university level in China.

II. Corpus in Language Learning and Translation Teaching

Just at the same time when Noam Chomsky made his impact on modern language studies in the 1960s, advocating the generative power of rules, other linguists such as Randolph Quirk became more and more aware of the inadequacy of a linguistic theory characterized by introspection alone. Real language data were needed for a more adequate and accurate description of language features. Started in the late 1950s, Quirk's *Survey of English Usage* is a collection of language data for empirical grammatical research, which led to one of the first corpus of English that became the source for studying standard English grammar for many decades.

Such real language data are of special use in foreign language teaching. A corpus of naturally occurring language data tells us what language is like, and it is a more reliable guide to language use than native speaker intuition is. Hunston exemplifies this by saying that "native-speaker language teachers are often unable to say why a particular phrasing is to be preferred in a particular context to another, and the consequent rather lame rationale 'it just sounds better' is a source of irritation to learners" (Hunston, 2002, p. 20).

With the advance in computer technology, the collected language data can be computerized into an electronic corpus. The electronic corpus is stored in such a way that it can be retrieved by corpus access software. The advantage of the data retrieval program is that it can automatically present all the instances that meet the search condition(s), which can best bypass arbitrariness and partiality of any human mind. Being nothing but a store of used language, a corpus in itself does not offer new information about language, but the corpus access programs such as those presenting concordance lines and calculating frequencies can serve as good tools to facilitate the researcher. Gerbig summarizes the possible uses of corpus in applied linguistics (Gerbig, 1997, pp. 43-44):

For lexicology, corpus provides inexhaustible instances and statistical evidence for the usage of words in actual contexts.

For syntax, corpus is the only way to conduct a quantitative probability survey of language structures.

For stylistics, general and specific corpora can offer a wide range of samples and background data for contrastive study of linguistic features.

Leech is even more confident in computerized corpus linguistics, regarding it as a new philosophical approach to the study of language. To him corpus linguistics (Leech, 1992, qtd. in Thomas and Short, 2001, p. 12)

…defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject. The computer, as a uniquely powerful technological tool, has made this new kind of linguistics possible. So technology here (as for centuries in natural science) has taken a more important role than that of supporting and facilitating research: I see it as an essential means to a new kind of knowledge, and as an 'open sesame' to a new way of thinking about language.

Short et al (2001, p. 112) highlighted the advantages of the corpus approach in that it "enables us to test out hypotheses in an explicit, empirical way, and to quantify the presence of categories and patterns of categories across text-types." They went on to point out the objectivity of such a machine-readable corpus approach, saying that it "forces the analyst to label every single part of a text and not to ignore examples inconvenient to the theory." (ibid) The last point is the very reason why the empirical approach outweighs the traditional intuition-based method; the repeatability and verifiability of the outcome is the true essence of any scientific research. The corpus approach is, therefore, a step forward towards the elimination of possible partiality and prejudice in selecting data in linguistic analysis.

Since the 1990s, corpus has found an increasingly important application in translation studies. Because corpora can enhance translation learner's awareness about language and culture, they are very "useful in training translators and in pointing up potential problems for translation." (Hunston, 2002, p.123). Bernardini (2003) suggests that by learning to effectively retrieve data from large-scale corpora, students majoring in translation can develop their translation skills effectively. Through numerous examples she has proved that corpora are able to help students understand the source text and translate it into a more readable target language.

III. Corpus Design for the Translator Training Course

Both comparable and parallel corpora are useful in student translator training. Comparable corpora contain two or more corpora in different languages or in different varieties of the same language. This means that a comparable corpus may collect texts in a particular field, but they are not actually translated from one another. They are collected because of their common text types and, therefore, their possible usefulness in facilitating translators and learners to identify differences and equivalences in each language. A comparable corpus of different varieties of the same language is actually a monolingual corpus. Such a corpus is useful in revealing whether and how translational language is different from original language, providing an empirical basis for the study of the often criticized "translationese". Parallel corpora usually comprise two or more corpora in different languages, each containing texts that have been translated from one language into the other. "Parallel corpora can be used by translators and by learners to find potential equivalent expressions in each language and to investigate differences between languages." (Hunston, 2002, p. 15).

There are some issues for consideration in designing corpora for student translator training.

a) Corpus design

With advanced technology in computer science since the 1960s, it is possible to store and access corpora of ever-increasing size. Finished in the 1990s, The British National Corpus (BNC) has largely expanded its size to 100 million words, and the Corpus of Contemporary American English (COCA), the largest corpus of English up to now, contains more than 400 million words. These two large-sized corpora are on-line accessible for free use, therefore they can be used by teachers to serve as reference for students to use for checking idiomatic English expressions in their translation learning processes.

Besides BNC and COCA, the translation teacher has to build other materials into the comparable corpora, depending on what particular purpose the translation class is aimed. In the university where I am now working, the English majors in the beginning of the third year are streamlined into two directions: philology and commerce, with different curriculum. We therefore decide to include not only literary texts but also texts of English for special purposes (ESP) in our corpora. For the sake of balance the literary corpora should include both modern and classic English literary works and the ESP should be as all-inclusive as possible: forensic, journalistic, clinic, economy, transportation and the like.

The design of the parallel corpora is more complex than the comparable corpora as it involves a translation relationship between two languages. The parallel corpora can comprise two parts, one is the reference corpora, which contain translations done by professionals, and the other is the learner corpora, which are composed of the assignments done by the students. The professional translations can again be divided into two, one is closely related to the classroom teaching and home assignment, providing recommendable versions to student assignment and it surely is more desirable if more than one version are provided for the same source text so that students can compare and contrast; and the other part is more comprehensive, providing a vast number and variety of translated texts with their source languages, covering all type of genres that the students are assigned to translate, so that students can pick up knowledge of style or a particular expression from the sea of information.

The learner corpora consist of the translation work done by the students. Each student is tagged with information about his or her age, sex, linguistic performance or other relevant data. What is especially interesting to the teacher as well as translation scholars are the progress the student makes in the course of learning with a time span of a semester, an academic year or even longer. This can be easily accessible with the electronically stored data. Such a diachronic study on student progress will also be encouragement to the students themselves.

Granger (2007, p. 21) outlines a figure of corpora in cross-linguistic research.

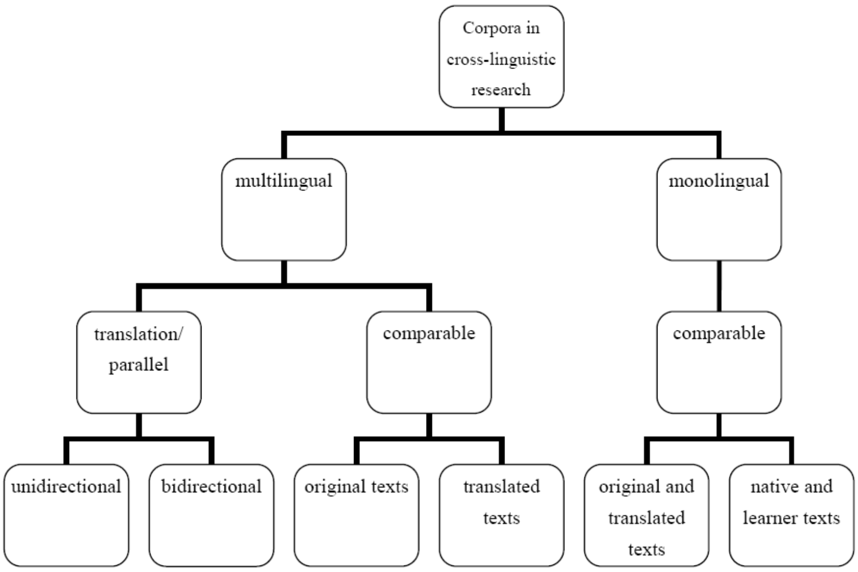

Figure 1 Corpora in cross-linguistic research

b) Sentence alignment

To align sentences in the parallel corpus, the initial step is to manually align the paragraphs in the source and target texts. Then sentence alignment can be done automatically with a software based on sentence or word boundaries or lengths, lists of anchor words, a bilingual lexicon, probability score, or a combination of algorithms. Mismatching sometimes occurs since automatic alignment may have difficulty in coping with restructured sentence orders in the translated text. A “txt” format of the aligned texts can make it more convenient to edit and proofread. All these correcting work must be done manually and can be very time-consuming, but it is a necessary procedure to ensure statistic accuracy as well as quick and exact retrieval of data for analysis. Sentence alignment is important in parallel corpora processing. Once a parallel corpus is aligned, a parallel concordancer can be employed to produce instances of occurrence of a word or structure in the source text and its equivalents in translation, or vice versa, thus valuable information can be extracted for translation studies. For example, how is a word in the source text translated in the target text under different contexts? How is the sentence structure changed after translation? Is there any recurrent pattern to be found? In addition, the information about the corresponding sentence ratio of the translated and the original provided by the sentence alignment software can also be very instrumental in a study of translation units and translation strategies such as combination and division at the sentence level.

c) Corpus annotation

The texts in the corpus are annotated with a header that can provide extra-textual information. The header includes the title, author/translator, language, text type, field, style, mode, time of publication, publisher and size of text. Besides header, annotation about the linguistic features of the texts is generally regarded to be more useful because annotation “adds value to a corpus, making it easier to retrieve information and increasing the range of investigations” (Hunston, 2002, p. 80). Automatic annotation software such as CLAWS has been developed to tag the part of speech of a word (POS tagging). Some researchers, however, are not keen on imposing POS labels on texts. One reason is that POS tagging has to be based on a specific grammatical theory, which is by no means agreed upon among all linguists. Experiments with manual checking of automatic tagging shows 2 percent disagreement among a group of linguists (Kennedy, 1998, p. 221), making the results and conclusions of different researches incomparable to some extent. Some researchers choose not to tag their corpora simply because their research does not require automatic linguistic differentiation.

d) Data-retrieval tools

The most frequently used tools in corpus-based learning is Wordsmith Tools. WordSmith Tools is a toolkit developed by Mike Scott at Oxford University. An integrated suite of programs for looking at how words behave in texts, its main functions include (1) the concord, which is a program that makes a concordance of a specified search word in the text file(s); (2) wordlist, which is a program that automatically generates word lists based on one or more ASCII or ANSI text files; and (3) the keywords, whose purpose is to locate and identify key words in a given text by comparing the words in the text with a reference set of words usually taken from a large corpus of text.

The "concord" function is basic in a corpus study, so the most common tool for data extraction is the concordancer. Using the concordancer, the user enters a search word, a "node", and the software will find all instances of the node in the corpus. Each instance is displayed with its immediate co-text and its filename, with the node highlighted and centered for prominence. All the instances can be sorted alphabetically or reverse alphabetically to the left or to the right of the node. This sorting function is very useful in that it allows the user to uncover possible patterns in all of the instances, and discard irrelevant instances for a particular study. The wordlist function is a frequency list of all the word in the corpus. Ordered either by frequency or by alphabet, it can be used to study word type, identify word clusters, compare the frequency of a word in different texts, compare translation equivalents between different languages, or to get a concordance of the words in the list. Kenny (2001) uses the wordlist function to analyze lexical items that occur only once, the "hapaxes", in her corpus in order to study lexical creativity in translation. A word or phrase with an extremely high (or low) frequency is worthy of scholars' attention because the frequency most convincingly reveals information about the idiosyncrasy of the text producer. The "keyword" function of WordSmith Tools is used to produce a keyword list, which is an extension to the frequency list. This is done by comparing frequency lists of two corpora with the help of the software. For example, by comparing the list from a large, general corpus with that of a smaller specialized corpus, the keywords of the smaller corpus can be identified.

IV. EVALUATION OF CORPUS-DRIVEN LEARNING IN TRANSLATION

Why do learners, after eight or more years of instruction in English, continue to feel headache in doing translation from Chinese into English? Apart from the vast discrepancy between the two languages, other possible factors may be attributed to improper teaching methodology. Corpus-driven learning is a recently developed methodology that is recommendable in its effectiveness to enhance students' language awareness and facilitate their lexical choice, syntactic variety and stylistic appropriateness in translation. When students are given access to more genuine language materials relevant to their current focus of study, they are endowed with more resources so that they can make better choices than when they have merely a few Chinese-English dictionaries at hand, which often fail to serve purpose. Since most entries in the bilingual dictionary are single words, and for most words we find many alternatives for how to translate them, but in most cases, the dictionaries can not tell us which of the alternatives we have to choose in a particular case, let alone what proper collocation that chosen word usually goes with. That is the reason why Teubert and Čermáková (2009, pp. 114-115) claim that bilingual dictionaries are not very helpful when the target language is not our native language.

That is where corpus, both comparable and parallel, comes into use. Just as Osborne (2009, p. 259) declares, "A major advantage of using corpus data in language learning is the possibility of making regularities in the language immediately more salient, by collecting dispersed naturally-occurring examples together as concordance lines, or by using these examples as a basis for language awareness exercises." Granger and Tribble (1998, pp. 199-209) also maintain that a corpus "can contribute to a better understanding of students' use of the foreign language".

In corpus-driven learning students can learn with the help of the data retrieval software, a concordancer, for example, which depicts frequent lexical/grammatical patterns of language within authentic contexts, presenting all the search results from huge amounts of linguistic data. Johns and King (1991, p. iii) describes data-driven learning as "the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language, and the development of activities and exercises based on concordance output." Batstone (1995) claims that data-driven learning is a pedagogic continuum from product to process. It has the advantage of product approach since the specific aspects of language are presented to the learners by multiple exposures within contexts. Corpus-driven learning also promotes creativity and self-discovery among learners. In corpus-driven learning, learners are not seen simply as recipients of knowledge, but as researchers who become more and more aware of the regularity of the language under study. In such a teaching method teacher's role is to encourage the learners' search for truth and enjoy the fun of discovery learning.

V. CONCLUSION

The use of electronic corpora is widespread in linguistics and translation pedagogy. The most obvious advantage of corpus over traditional teaching methodology is that it enables students to have a quick access and analysis of large quantities of material. Through learning from such extensive material it will be possible for students to obtain more reliable information on translation norms and patterns. This is of special significance in translation education in China, given the limited credit hours of E-C and C-E translation in the curriculum design in some universities.

The application of corpus methodology in classroom translation teaching does not mean that the teachers can finally stay away, leaving students to sink or swim on their own. On the contrary, corpus-driven learning methodology poses even more challenges to the teachers. They must devote more time to building the most facilitating corpus for use, specifying learning tasks that can take the best advantage of the corpus, and, most importantly, to help students to arrive at a sound interpretation of the (large) data retrieved from the corpus. Good facilitator as the computer is, it can never totally replace the human role assumed by the teachers in the process of interactive learning, especially in the field of translation, which is, by nature, a very complex process of cross-lingual, cross-cultural communication.

REFERENCES

[1] Batstone, R. (1995). Product and process: Grammar in the second language classroom. In M. Bygate, A. Tonkyn, and E. William, (Eds.), *Grammar and the language teacher* (pp. 224-236). London: Prentice Hall.

[2] Bernardini, S. (2003). Corpora in Translator Education: An Introduction. In F. Zanettin, et al. (eds.) *Corpora in Translator Education*. London: St. Jerome.

[3] Gerbig, A. (1997). Lexical and grammatical variation in a corpus: A computer-assisted study of discourse on the environment. Frankfurt and Berlin: Peter Lang GmbH.

[4] Granger, S. et al. (2007). Corpus-based Approach to Contrastive Linguistics and Translation Studies. Beijing: Foreign Language Teaching and Research Press.

[5] Granger S. and C. Tribble. (1998). Learner Corpus data in the foreign language classroom: From focused instruction and data-driven learning, in S. Granger (ed), *Learner English on computer* (pp. 199-209), London: Longman.

[6] Hunston, S. (2002). Corpus in Applied Linguistics. Cambridge: Cambridge University Press.

[7] Johns, T. F. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. F. Johns and P. King (Eds.), *Classroom concordancing* (pp. 27-45). Birmingham: ELR.

[8] Kennedy, G. (1998). An Introduction to Corpus Linguistics. London: Longman.

[9] Kenny, D. (2001). Lexis and creativity in translation. A corpus-based study. Manchester: St. Jerome.

[10] Oldin, T. (1989). Language transfer: Cross-linguistic influence in language learning. Cambridge: Cambridge University Press.

[11] Osborne, J. (2009). Top-down and Bottom-up Approaches to Corpora in Language Teaching. In U. Connor and T. A. Upton (eds.). *Applied Corpus Linguistics: A Multidimensional Perspective* (pp. 251-265). Beijing: World Publishing Corporation.

[12] Teubert, W. and A. Čermáková(2009). Corpus Linguistics: A Short Introduction. Beijing: World Publishing Corporation.

[13] Thomas, J. and M. Short. (2001). Using Corpora for Language Research. Beijing: Foreign Language teaching and Research Press.

[14] Towel, R. and R. Hawkins. (1994). Approaches to second language acquisition. Cleveland: Multilingual Matters.

Qing Wang was born in Yantai, China in 1969. She graduated from Shandong University, China, in 2010 with Ph. D in Translation Studies. She is currently associate professor in Department of English, Shandong Jiaotong University, Jinan, China. Her research interests include translational stylistics, literary stylistics, and corpus translation studies. Dr. Wang is a member of China Association for the Philosophy of Language.