A Corpus Pioneer in Chinese-English Lexicography

Hugo T. Y. Tseng
Soochow University

The Chuan-Shi Chinese-English English-Chinese Thesaurus (Liu 2005, henceforth Chuan-Shi) and its earlier editions attempt to redress such inadequacies by starting with the collection of a written corpus of over 10,000 English sentences or paragraphs closely related to Chinese culture. This corpus was then used to extract some 65,000 Chinese-English interlingual equivalents, which can be looked up using either the Chinese or English index. The originality shown in Chuan-Shi has demonstrated the vital but long-neglected role which computer corpora may play in Chinese-English lexicography (Tseng 2001, 2004). Liu was a translator and copy-editor of Chinese and English publications for 30 years, after which he spent another 10 years on the compilation of this volume to resolve the difficulties he himself had encountered during the course of his work. Such an innovative approach could not only shed new light on Chinese-English lexicography, but also inspire bilingual lexicography in general.

Key words: Chuan-Shi, lexicography, bilingual dictionary, corpus driven

1. Introduction

Current Chinese-English dictionaries are inadequate in several ways. First, some of the frequently sought-after, culture-specific expressions in daily use are missing, inadvertently or purposely. Second, in cases where they are entered, the English equivalents offered are usually unidiomatic or even inaccurate, and many of these “equivalents” are just paraphrases or lengthy explanations at most. The absence of truly text-insertable translation equivalents is hardly helpful in translating from Chinese to English.

The Chuan-Shi Chinese-English English-Chinese Thesaurus (Liu 2005, henceforth Chuan-Shi) and its earlier editions attempt to redress such inadequacies by starting with the collection of a written corpus of over 10,000 English sentences or paragraphs

* This article was originally an unpublished paper (“Constructing a Corpus for Chinese-English Lexicography”) presented at the third ASIALEX Biennial International Conference in Japan, August 27-29, 2003. The current version has been revised extensively.
closely related to Chinese culture. This corpus was then used to extract some 65,000 Chinese-English interlingual equivalents, which can be looked up using either the Chinese or English index. The originality shown in *Chuan-Shi* has demonstrated the vital but long-neglected role which computer corpora may play in Chinese-English lexicography (Tseng 2001, 2004). Liu was a translator and copy-editor of Chinese and English publications for 30 years, after which he spent another 10 years on the compilation of this volume to resolve the difficulties he himself had encountered during the course of his work. Such an innovative approach could not only shed new light on Chinese-English lexicography, but also inspire bilingual lexicography in general.

2. Contemporary Chinese-English lexicography: a quick overview

Two epoch-making dictionaries, *A New Practical Chinese-English Dictionary* (Liang 1971) and *Lin Yutang’s Chinese-English Dictionary of Modern Usage* (Lin 1972), had dominated the scene outside the PRC for nearly a decade, mainly because of the intellectual hiatus caused by the Cultural Revolution. Revisions of Liang’s and Lin’s dictionaries appeared in 1992 and 1987 respectively under different titles by different authors, but these two lines of products essentially reflect the standard Chinese language from a more traditional, non-PRC perspective.

In the PRC, Chinese-English lexicography has flourished for a generation since the appearance of the landmark monolingual *Xiandai Hanyu Cidian (A Contemporary Chinese Dictionary)* in 1978. This medium-sized prescriptive dictionary served as a source language blueprint and sparked the publication of *A Chinese-English Dictionary* (Wu 1978), from which a multitude of other Chinese-English dictionaries mushroomed in the ensuing decade. The next decade of 1990’s witnessed the emergence and growth of large-sized dictionaries. In 1993, the *Chinese-English Dictionary* in two big volumes (edited by Wu Guanghua) was published by Shanghai Jiaotong University Press, and was awarded Chinese Book Prize the following year. In 1995, the updated and revised edition of Wu (1978) was published by the Foreign Language Teaching and Research Press under the chief editorship of Wei Dongya, and won the first award in the 1997 biennial Chinese Lexicography Prize. In 2000, the Commercial Press produced the *New Age Chinese-English Dictionary* (co-edited by Wu Jingrong and Cheng Zhenqiu). It has since received considerable critical acclaim and earned the first award in the 2001 biennial Chinese Lexicography Prize. In late 2003, Hui Yu’s *A New Century Chinese-English Dictionary* came out under the imprint of the Foreign Language Teaching and Research Press. The next salient tour de force in Chinese-English lexicography, *A Great Comprehensive Chinese-English Dictionary*, came out in early 2004 in three big volumes. Again edited by Wu Guanghua, this dictionary claims to be the largest ever. Its physical
size, lexical coverage, and number of definitions have surpassed all its predecessors.

Outside the Chinese-speaking world, the most conspicuous effort came from the *ABC Chinese-English Comprehensive Dictionary* (1996, 2003). Chief editor John DeFrancis is one of the most revered contemporary American linguists of Chinese. Macrostructurally, he placed sound (Pinyin) before form (of Chinese characters), adopting a strict, one-stop alphabetical Pinyin macrostructure. Microstructurally, he pioneered some morphological, pragmatic, and syntactic analyses of the 200,000 headwords.

In spite of all the prosperity on the Chinese-English lexicographical arena, there have been few innovations in the way these dictionaries were produced. With the only possible exception of DeFrancis (1996, 2003), the macrostructures and microstructures essentially remain the same. Head characters are alphabetically arranged according to Pinyin, the official romanization system for transliterating Chinese. Right under each head character, translations and/or explanations are given with possible illustrative examples, and further sub-entries of character combinations, left-justified in Pinyin order, are similarly treated. What these dictionaries have been competing for were little more than meeting some general expectations that make a good dictionary—lexical coverage, the inclusion of new words and senses, the number of illustrative examples, and the choice of equivalents or the wording of explanations.

The core task of a bilingual dictionary is to provide users with interlingual equivalents, which is an extremely tough job. The aforementioned award-winning Chinese-English dictionaries claim the accuracy and reliability of their English translations and/or explanations. However, the validity of their claim is open to question. The blurbs on the back covers and the prefatory remarks inside reveal that the entries and sub-entries have been examined and approved by local and/or native English experts. In the prefatory matters, Wu (1993, 2004) are reservedly candid in admitting that the contents do not solely rely on any particular dictionary, but are distilled from the best of all dictionaries. I take such statements as an euphemism for collating equivalents, explanations, and examples from multifarious reference books and packing them into his own work.

Apparently, the editors of these dictionaries still rely on personal introspection. They resort to their intuition, judging the validity of linguistic data by their linguistic competence. This approach has been widely adopted, and it does save time and money. However, even if the editors are recognized experts in the English language, they may be constrained by the scope of their knowledge, experience, memory, and preference. These experts in English may produce such run-on sentences as *After all he is a foreigner, he cannot use chopsticks well.* Likewise, sentence fragments may also be found in these experts’ writings: *Whenever I talk to my adviser about my thesis.* Quasi-English expressions, such as *morning call* (for *wake-up call*) or *DM* (for *leaflet*), may be produced, or considered acceptable, by these experts. If these experts don’t
drink coffee, *cappuccino* and *latte* may be totally incomprehensible to them. If they have little knowledge about Western food, *quiche* and *lasagna* will not be considered common vocabulary items. It is doubtful whether these “experts in English,” when hired as editors or consultants of Chinese-English dictionaries, can perform a decent job in making linguistic judgments, although such an introspective approach has been practiced since time immemorial.

3. A Corpus for Chinese-English lexicography

Since 1980 when COBUILD (Collins Birmingham University International Language Database) started to run, the corpus has played a vital role in English lexicography. The traditional introspective approach has been challenged. This huge corpus, later known as the Bank of English, currently contains over 650 million words of running text. Its application in linguistic analyses and lexicography is revolutionary, so other multinational publishers, including OUP, Longman, CUP, and Macmillan, has all jumped on the corpus bandwagon.

By contrast, the application of corpora to the compilation of bilingual dictionaries with Chinese and English has lagged considerably. The scope of such application, if any, is much smaller. But the use of corpora in the compilation of dictionaries appears to be irresistible. The English-Chinese dictionaries have gradually sensed the trend, and started to gather their own corpora. However, the Chinese-English dictionaries, except for *Chuan-Shi*, have not made noticeable progress in this regard.

The most remarkable innovation of *Chuan-Shi* is to first systematically apply the corpus the author has collected to the compilation of a Chinese-English dictionary. Although the sources are limited in number, genre, and variety, this book is the fruit of painstaking labor of a dedicated individual without any support from academia, publishers, or governments. As far as the macrostructure is concerned, *Chuan-Shi* differs markedly from other Chinese-English dictionaries in that it is more like a thesaurus of citations, which are categorized under 1,281 head concepts and further to subordinate concepts. As far as the microstructure is concerned, the citations are grouped together with key expressions in bold-type font and immediately next to each, glossed in Chinese. Although looking up specific words is quite complex, the methodological innovations in *Chuan-Shi* presents a new form compared to the traditional Chinese-English lexicography. *Chuan-Shi*’s citation files, apparently inspired by the OED’s Reading Program, contain English sentences and paragraphs on Chinese culture, from which authentic key expressions are extracted and translated back to Chinese. Take for instance a four-character idiom *hutoushewei* “tiger’s head and snake’s tail.” Most Chinese-English dictionaries render it as “fine start and poor finish” or “do things by halves.” But in stark contrast, *Chuan-Shi*
does the following:

When the verdict finally came…it was almost an anteclimax 虎頭蛇尾. *Fox Butterfield, China Alive in the Bitter Sea, 1982, p. 359.*

By doing so, the author reverses the traditional lexicographical process of starting with the choice of Chinese headwords and ending with providing translations/explanations and illustrative examples. He gives the users authentic, idiomatic, corpus-evidenced, and text insertable equivalents, a goal which other lexicographers can hardly achieve by introspection.

Chinese-English dictionaries for Chinese speakers should be geared to encoding tasks, that is, helping dictionary users in expressing in the “code” of the English language what they have in mind in Chinese “code.” Dictionary users do not want many explanatory equivalents; rather, they are almost always in need of idiomatic translation equivalents. On the other hand, of course, Chinese-English dictionaries for non-Chinese should give more help in “decoding” the Chinese expressions for the foreign language learner. Their need for explanatory equivalents is supposedly higher, because explanations facilitate comprehension. *Chuan-Shi’s* systematic extraction of authentic interlingual equivalents and illustrative examples from the corpus is a remarkable step forward in Chinese-English lexicography. The equivalents and illustrative examples thus culled are naturally more idiomatic and readily text insertable, meeting the requirements of encoding tasks to be performed by native Chinese users.

4. Corpus-driven—Future trend of Chinese-English lexicography

Unlike *Chuan-Shi*, however, Chinese-English dictionaries currently on the market tend to be ambivalent about their targeted group of users, or claim to serve both native and non-native Chinese speakers. In spite of their claims, the truth is that these dictionaries prioritize Chinese-speaking users, but the commercial strategy of killing two birds with one stone fails to please any party. These Chinese-English dictionaries resort to introspective explanations when the equivalents of culture-specific expressions are hard to come by. Lengthy and not text insertable, these explanations do not help language production—the most important task Chinese users wish to perform. Also, because the translation equivalents and illustrative examples provided in these traditional Chinese-English dictionaries are not corpus-based, but rather the product of editors’ introspection, their authenticity and idiomaticity are open to question.

The innovative approach adopted by *Chuan-Shi* can be incorporated into the design of traditional Chinese-English dictionaries targeting Chinese users. One of the most
important strengths of these Chinese-English dictionaries is their clear macrostructure and easiness of information retrieval, both of which should stay. Their most prominent weakness, in comparison with *Chuan-Shi*, is the absence of corpus evidence. The optimal future for the Chinese-English dictionaries for the Chinese is an eclectic union of Pinyin macrostructure and corpus-based translation equivalents and illustrative examples.

Though building a bilingual corpus comparable in size and sophistication to the Bank of English is still a dream, it should be regarded as a must. The age of corpora has arrived, and we should do our utmost to catch up. At the early stage of corpus building, a team could gather English-language newspapers, magazines, books, Web pages, etc. that touch on Chinese culture and society, and have them stored electronically. After that, a bilingual reading program starts. A large number of Chinese-speaking people competent in English scan the raw materials, highlight the key expressions, translate them into idiomatic Chinese, and have them tagged for future analyses and retrieval. At a later stage, this corpus should be further expanded to include general, in addition to the Chinese culture-specific, materials. Only by this way can a fuller picture of interlingual equivalence be presented for a more detailed and comprehensive treatment.

The reason why the building of a bilingual corpus should proceed by stages is closely linked with the nature of interlingual equivalence. Snell-Hornby (1990:210) distinguishes five levels of interlingual equivalence, which are, from the simplest to the most complex, (1) terminology/nomenclature, (2) internationally known items and sets, (3) concrete objects, basic activities, stative adjectives, (4) words expressing perception and evaluation, often linked to sociocultural norms, and (5) culture-bound elements. This categorization is applicable to the prioritization of the building of a bilingual corpus, the emphasis of the reading program and the inclusion of equivalents and illustrative examples. The more complex the interlingual equivalence, the more urgent is such a program, and the more equivalents and examples should be included. The interlingual equivalence of the first two categories is simple and straightforward, so a valuable space could be reserved for the last two categories.

*Chuan-Shi’s* application of a corpus points the way to the future of Chinese-English lexicography. Macrostructurally, such a prototypical dictionary enters its head characters alphabetically according to Pinyin, and under each head character, a certain number of Pinyin-ordered character-compound sub-entries are headed by the respective head character. All their English equivalents and illustrative examples are extracted from a huge bilingual corpus, which ensures that awkward translation equivalents and lengthy explanatory equivalents are eliminated. For possible entries and sub-entries not supported by the corpus, the corpus should eventually be expanded. But before this is complete, lexicographers should still have recourse to the old approach—making up translation equivalents for text insertion and supplement them with explanatory equivalents.
References

Dictionaries


Non-dictionary Sources