Lexicography Across Languages

YUKIO TONO

Most lexicographical theory and practice is based on the analysis of a first language (L1), but lexicography can be extended to the realm of "across languages," and is then called bilingual or multilingual lexicography. It is an area of particular relevance for language learners and translators. This entry will focus on a particular aspect of multilingual lexicography: how technology has affected the innovations of dictionary making involving more than one language. (For a more general introduction on bilingual dictionaries, see Svensén, 2009.) I will begin by briefly reviewing some basic concepts and then discuss technological advances and related issues in lexicography across languages. Finally, some recent innovations in bilingual lexicography will be described (with a special emphasis on the use of corpora) in relation to second language (L2) vocabulary learning studies and user needs/skills research.

Basic Concepts of Lexicography Across Languages

Before we can say much about the technology of lexicography across languages, there are a few concepts and terms that we need to clarify. Bilingual dictionaries specify, for words and expressions in the *source* or native language (L1), their equivalents in the *target* language, that is, another language that is not native for the user (L2). As in the case of monolingual dictionaries, these dictionaries can be used in two directionally opposite ways: *encoding* and *decoding*. Encoding is the process of putting a sequence of words into a sentence (or a paragraph or a discourse) in order to convey meanings. Decoding is the opposite process. Therefore, there are two different types of bilingual dictionary:

- an L2–L1 *reception/translation* dictionary (for understanding L2 text), and
- an L1–L2 *production/translation* dictionary (for producing L2 text).

Mono- or *unidirectional* dictionaries usually contain a single text in which the target language is explained in the source language; but some dictionaries have two sets of texts and work both ways (i.e., L2–L1 and L1–L2), and are called *bidirectional* in the sense that both groups of language speakers, whether L1 or L2, can use them in either direction for either purpose (encoding or decoding).

Another important distinction is between *bilingual* and *bilingualized* dictionaries. Bilingual lexicography traditionally involves the translation of monolingual dictionaries (or databases). Thus it is common for major dictionary publishers to have various bilingual versions of their monolingual dictionaries published in foreign countries. In such cases, they usually provide a database or framework for the target language dictionary entries, which contain all the information relevant to the creation of a dictionary in the target language (i.e., monolingual format), and local publishers will only add translations (the process technically called "transfer"; Atkins & Rundell, 2008, p. 465) to the necessary entries. In countries where bilingual lexicography is relatively slow in development, so-called *bilingualized* dictionaries are popular; these are basically monolingual dictionaries in their original

The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle.

DOI: 10.1002/9781405198431.wbeal0700

format with minimum semantic glosses in the native language alongside each definition. In countries where bilingual lexicography is more fully developed, mere translation of monolingual dictionaries is not appreciated, and a complete bilingual framework will be used in order to produce original bilingual dictionaries.

There are some issues specific to bilingual lexicography. For example, the transfer between monolingual and bilingual entries involves problems of mapping lexical items between the two languages. We try to work out a balance between the "reality" as it is perceived in the two cultures and the words in the languages which describe it (Duval, 2008). Total equivalence cannot be achieved in most cases, and lexicographers resort to a near-equivalent or a gloss to fill the semantic gap of direct translations. In other cases, it is also possible to supplement the direct translation by using usage notes. All these devices need careful planning and guidance for lexicographers. This practice of translation "transfers" can also be applied to grammar, labels, and illustrative examples.

While major monolingual learners' dictionaries (e.g., the *Longman Dictionary of Contemporary English*) already draw attention to frequency variation in speech and writing, bilingual dictionaries have a further role: They have to deal with the problem of choosing the right translation equivalents whose frequencies are almost equal to the source language items in the same registers. This is often a difficult task, but would be very important for those who create subtitles for movies or TV dramas, or for translators of speeches made by government ministers and other VIPs. In practice, bilingual lexicographers sometimes deliberately ignore this aspect of variation, especially those who work with learners' dictionaries. Teachers are still conservative and prefer formal translation equivalents to informal, impolite expressions. They often think it is not appropriate to list offensive words in the entries.

Technological Advances and Issues in Lexicography Across Languages

While many bilingual dictionary projects are still based on translation within a monolingual framework, there are some interesting technological advances that are changing the way bilingual/multilingual dictionaries are being made. Three of them will be discussed here; (a) bilingual lexicon acquisition from parallel corpora; (b) the integration of parallel text databases into electronic dictionaries; and (c) multilingual, multimodal dictionary portals.

Semi-Automatic Identification of a Parallel Terminology

The information explosion of the Internet has made bilingual texts increasingly available. It is now possible to use translation pairs found in such texts for the automatic compilation of a bilingual lexicon. Matsumoto and Utsuro (2000) give a basic overview of this process: A pair of bilingual texts, which are translations of each other, is prepared. The texts will be *aligned* in the sense that there is an order-preserving mapping between sentences in parallel texts. Then the texts will go through a series of preprocessing steps, including (a) lemmatization/segmentation (mapping surface word forms to the underlying base forms), (b) part-of-speech (POS) tagging (assigning word class tags to each word), (c) shallow-parsing/noun phrase (NP) recognition (identifying and marking NP chunks), and (d) parsing/bracketing (analyzing syntactic or dependency structures). The processed bilingual texts will then be compared and analyzed by computer, using various heuristics such as machine readable dictionary (MRD), cognate, POS, or positional heuristics, in order to determine translation pairs. For instance, it is more likely that two sentences are translations of each other if they share the same linguistic features, such as lexical items identified as translation equivalents shown in MRD. Translation pairs could be simply single words, but may also be larger units such as NOs, collocations, and dependency structures (i.e., structures showing dependency relations between the words in a sentence). So far, most of the attempts at bilingual lexicon construction using parallel texts have concentrated on the correspondence between words or NPs, and significant progress has been made in the extraction of word-/NP-level translation lexicons (see Matsumoto & Utsuro, 2000, for further detail). There are already many such bilingual terminological dictionaries semi-automatically created using the technologies described above (e.g., the *Japanese–English Dictionary of Technical Terms*; CJK Dictionary Institute, 2009).

Integration of Parallel Text Databases Into Electronic Dictionaries

Major online dictionaries (e.g., The Free Dictionary.com, www.thefreedictionary.com, or Dictionary.com, http://dictionary.reference.com) are usually based on unabridged monolingual dictionaries such as the *Random House English Dictionary* or the *American Heritage English Dictionary.* If bilingual functions are available, in many cases (e.g., FREELANG Dictionary, www.freelang.net) they provide only bidirectional word lists and little other information. In this respect, Eijiro (for the Web version, see www.alc.co.jp) is unique. It has developed a hybrid resource that combines ordinary electronic dictionaries with parallel text databases. Its entry structure, too, is unique: Instead of having a traditional dictionary entry format, each piece of information, whether it is a word, phrase, or example, has independent entry status, and everything is in parallel in English and Japanese. Thousands of hard-to-translate phrases, proper names, and compounds are collected and put into the database. As the number of entries has reached 1.8 million, Eijiro can serve as both a bilingual dictionary and a bilingual parallel text database, which plugs the gap in collocational information that exists in current bilingual dictionaries.

Multilingual, Multimodal Dictionary Portals

The advent of Internet technologies has improved the design of online dictionaries radically. Lexicool.com (www.lexicool.com), a directory of online bilingual and multilingual dictionaries, has links to over 7,000 dictionaries and glossaries. The English dictionary section has more than 60 languages as target languages and each language page has further links to hundreds of dictionaries (e.g., 1,201 for German, 1,799 for French, and 197 for Japanese, as of February 2010). Recent trends are the crossover of dictionary entries and various other contents related to the search words. For example, WordReference.com (www.wordreference.com) has more than 10 different bilingual dictionaries (English–Spanish/French/Italian/German/Russian/Polish/ Romanian/Turkish/Chinese/Japanese/Korean/Arabic, etc.), many of which can be searched bidirectionally. If the English word *flower* is searched in English–French dictionaries, the results show several different types of lexicographic information: (a) the entry from the *Concise Oxford-Hachette English–French Dictionary*, (b) the other headwords that contain *flower* in the definitions, (c) example sentences from English–French translation pairs that contain the word *flower*, and (d) compound forms (e.g., *basket flower*, *flower arrangement*, *flower show*, etc.). The site searches information across different dictionary sets and provides summaries of the output.

The very notion of what a dictionary is has been changing since the advent of the Internet. Dictionary portals such as WordReference.com expand their offerings by incorporating various online resources and information related to the search words that users key in. The advantages of the social networking potential of the Internet add even greater extra value to these sites. In addition to general lexicographic information for the search word, WordReference.com provides links to Internet forums among translators, in which people discuss how to translate various expressions related to the search word in both the source and target language. The discussions and comments in these forums can prove very helpful in solving problems that individual translators may be facing. In addition, WordReference.com links to many other languages, along with Google search results (see also http://translate.google.com), and images related to the search word.

Although bilingual lexicography has benefited from the technological advances described above, several issues remain to be resolved. First, although automatic extraction of bilingual lexicons is relatively easy for fixed technical terms, translation of literary texts or imaginative writings is still extremely difficult. Cultural, stylistic, and rhetorical factors affect the choice of appropriate translation equivalents in these domains, and it is still up to the skills of individual human translators or lexicographers to choose the right equivalents. Second, while online dictionaries such as Eijiro or WordReference.com are extremely useful, there is always a chance of information overload. Users can be overwhelmed by the mass of information returned from the query. Only skilled users can make full sense of the results, and the reference acts of less skilled users tend to be very limited and problematical, because they only look for the information shown at the very top of query results. Third, we should always observe carefully the balance between cost and quality. As Web resources grow rapidly and a growing number of open-ended free dictionary sites (e.g., Wiktionary, http://en.wiktionary.org) become available, there is a genuine concern among professional lexicographers that the quality of the contents of the free dictionaries is not as high as that of commercial ones. While Wikipedia offers "facts" about the world, Wiktionary offers "word meanings" of a word. Writing an encyclopedia entry requires profound knowledge about the subject in question, but any serious enthusiast or hobbyist may have sufficient knowledge to write an acceptable entry. Writing a dictionary entry, however, requires more professional expertise, as it takes years of training and experience in analyzing and creating word meanings to write good definitions. Wiktionary contributors are likely to have varying levels of such expertise.

Recent Innovations in Bilingual Learners' Dictionaries

While all approaches to bilingual lexicography are based on Atkins and Rundell's (2008) three steps (*analysis*, *transfer*, and *synthesis*), various new ideas have been introduced and tried out in an effort to make dictionaries more user-friendly. Here we will discuss three topics: (a) corpus-informed descriptions, (b) integration of vocabulary learning strategies, and (c) user needs/skills research.

Corpus-Informed Descriptions

Since the publication of the *COBUILD English Dictionary* in 1987, the use of corpora has been popular, especially in pedagogical lexicography. Major monolingual learners' dictionaries now provide detailed information about the frequency of the common words in their respective headword lists. Collocations or chunks taken from corpora are often shown as useful expressions. Example sentences are either selected from corpora or rewritten from the corpus evidence (Rundell, 1998). Cobb (2003) even suggested that actual concordance lines could be displayed in corpus-based electronic dictionaries. In Japan, a pocket electronic dictionary manufactured by SEIKO Instruments Inc. (SII) in 2004, for the first time in the world, made the *COBUILD English Dictionary* available with an additional 500-million-word "WordBank," the actual corpus data from the Bank of English.

For specific target learners, the use of learner corpora (corpora of L2 learners' speech or writing) is found to be effective (Tono, 2009). While some monolingual learners' dictionaries (e.g., the *Longman Dictionary of Contemporary English*, *Longman's Essential Activator*, the *Cambridge Advanced Learner's Dictionary*) have used learner corpora for their "common learner error" notes or columns, it has been found that specific learner groups are particularly prone to specific patterns of error. If the aim of a dictionary is to serve the needs of particular target user groups, it is desirable to incorporate the information on common errors made by learners of specific L1 backgrounds. As the *Cambridge Learner's Dictionary* was once semi-bilingualized in Japan, the editor on the Japanese team investigated the Japanese learners' subcorpora in the Cambridge Learner Corpus and added extra columns on learner errors specific to Japanese learners of English. Such localization is necessary in order to make bilingual versions more useful.

Second, there is a growing awareness that usage notes in pedagogical dictionaries should take into account the developmental aspects of L2 learner proficiency. For this, a corpus of L2 learners' speech or writing, called a learner corpus, is useful. One of the special features of learner corpora is "error tagging," which annotates different types of learner errors (Granger, 2003). By using error-tagged learner corpora, we can extract frequency information on different error types, which will contribute to the analysis of learner language and how to provide pedagogical support for overcoming errors. Most "common learner error" notes in monolingual dictionaries are based on the analysis of the writing of relatively advanced learners. An analysis of a corpus of L2 learners of specific L1 background (Tono, 2009) has shown that there is a relationship between particular error patterns and stages of acquisition. For example, verb morphology errors are more common for beginning and intermediate-stage learners, while lexical choice errors are observed more commonly in advanced learners' writing. Therefore, it is desirable to provide different types of "common error" notes for different levels of users. Findings based on the analysis of target-user L2 learner corpora should be an integral part of usage descriptions in bilingual learner dictionaries.

Another interesting application of corpora in bilingual lexicography is a corpus-based dictionary of synonyms (Tono, 2005). This dictionary contains about 200 Japanese entries, each of which have three to five English translation equivalents. For example, the Japanese word *kashikoi* can be translated as 'clever', 'intelligent', or 'wise'. The problem is that Japanese learners of English often find it difficult to use these English words in a proper way. Typical dictionaries of synonyms list all these and explain the differences in meanings with some illustrative examples. This kind of metalinguistic explanation is not always helpful, often giving learners the extra burden of having to understand the metalanguage. Tono (2005) took a totally opposite approach. By showing the most typical collocations for those three words, the dictionary enables users to compare the usages of similar words and figure out how to use them in different contexts. The following are the lists of five collocates for the three English adjectives, using a measure of strength of association called log-log scores, based on the British National Corpus:

clever + noun	*intelligent* + noun	*wise* + noun
1) girl	1) being	1) precaution
2) boy	2) man	2) man
3) trick	3) network	3) counsel
4) man	4) people	4) decision
5) idea	5) woman	5) woman

A comparison between the synonyms shows that the word *clever* tends to collocate with younger people (e.g., *girl* and *boy*) and that it also has some connotations of being slightly dishonest (e.g., *trick*). The word *intelligent*, on the other hand, can collocate with people in general who have a high level of mental ability and also systems that can work like human beings. The word *wise* typically collocates with actions such as *precaution* or *decision*, implying good, sensible judgments. Some overlaps among the collocates for the three English synonyms above indicate that there are some meanings shared across the three, whereas different collocations highlight semantic as well as usage differences among them. A list of collocation patterns alongside summary usage notes will greatly enhance learners' understanding of the usage of the synonyms.

Integration of L2 Vocabulary Learning Strategies

As research into L2 vocabulary learning processes and strategies increases, there is a growing awareness that insights from such research fields should enrich dictionary design. The results of corpus studies show that there are some distinct characteristics in the way vocabulary behaves in language use. For example, the most frequent 100 words will cover approximately 65–7% of spoken data, and in most texts around 80% or more of the words are from the most frequent 2,000 word families. In casual conversation, over 90% of the words tend to be from this range of words (Nation, 2008). Therefore, it would be useful for learners to be aware of the relative importance of the words that they encounter in a dictionary.

Editors of bilingual learners' dictionaries have begun integrating information on the degree of relative communicative utility of lexical items into their dictionary designs. In the *Ace Crown English–Japanese Dictionary* (Tono, 2008), for example, two-page special columns called "Focus Pages" are prepared for the most frequent 100 headwords, representing 25 major lexical verbs and function words (auxiliaries, prepositions, conjunctions, and *wh*-words). Each special column focuses on essential information about a headword, such as its core image (usually accompanied by illustrations), meaning maps (semantic relationship among the senses), corpus-based information such as the most salient syntactic patterns, collocation patterns, and some illustrative examples from English textbook corpora as well as native speaker corpora. While this kind of information has been available sporadically in monolingual learners' dictionaries, few highlight the features in a comprehensible manner in order to encourage users to acquire deep vocabulary knowledge for these entries. The *Ace Crown* takes special care to focus on verb patterns, polysemous entries, and collocation knowledge, which are all supplied with corpus frequency ranking information. Learners can visually see how much knowledge has to be acquired to have productive knowledge of the given headword.

The dictionary also shows clear distinctions between the most frequent 2,000 words and the next top 3,000 words, whose entries are marked in asterisks and in red. Additional information useful for productive purposes is supplied for the top 2,000 words, while the entries for the next 3,000 most frequent words show essential information for receptive purposes only. All the rest of the approximately 50,000 entries have Japanese translation equivalents only, which is meant to communicate to users that those are the words that they do not have to memorize or learn in depth, but can look up in a dictionary whenever they are encountered. This kind of distinction between active and passive vocabulary has not been clearly made in learners' dictionaries until recently, because dictionaries serve multiple purposes for multiple groups of users. However, if the target user group is clear, learners' dictionaries should support the vocabulary learning process of users, and specify exactly what vocabulary knowledge users need for that level, and for what purpose.

User Needs/Skills Research

User needs/skills research is also an important area in lexicography. Dictionaries are the primary (and sometimes only) tools that learners turn to to solve almost every language difficulty, and thus should ideally be a comprehensive, one-stop resource. In reality, however, it is impossible to predict all the questions that users will ask of their dictionary, so we need to take a pragmatic view: "A realistic goal is to meet the needs of most users most of the time" (Atkins & Rundell, 2008, p. 32). Recently, lexicography across languages has gone in two different directions. One is to aim at inclusiveness, providing "everything" that potential users might look for. Online dictionary sites like Eijiro or WordReference.com are such cases. They do not gear their contents toward specific types of users. Instead, they include as much information as possible, not only lexicographical but also encyclopedic and sociolinguistic, and it is up to the users to get what they want. Since there is virtually no space limit, owing to the use of digital technology, this trend will continue until somebody questions the value of such masses of information and restructures the information categories.

Another direction is to make dictionaries more and more specialized. This is again linked to the development of the Internet, and a most striking difference from 10 years ago is that nowadays more and more free terminological dictionary sites are available. The GLOSSARIST portal site (www.glossarist.com) has more than 20,000 glossaries and topical dictionaries (both monolingual and bilingual), many of which are developed by professionals or translators working in that particular field. The problem is that they are mainly compiled for professional reference or research purposes, and are usually neither systematic nor exhaustive. Thus, from a user perspective, there must be some kind of criteria for evaluating and filtering those sites. It might be useful to establish a consortium to recommend a list of evaluation criteria to guide the creation of such sites in order to standardize some of the information contents. Despite their varied nature, such specialized sites are updated constantly, and new terms are added quickly, so they are invaluable resources for those who know what to look for.

Conclusion

Increasing numbers and types of bilingual and multilingual dictionary have become widely available since the development of the Internet. It might be good to exploit the power of Internet social networking tools to allow everyone to get involved in dictionary making (see, for example, an early attempt by Cobb, 1999). However, as is the case with wiki lexicographers, this needs careful quality control. Having said that, in theory, the transfer of information across languages has inherent linguistic and cultural difficulties, which will never be perfectly solved even in the Internet age. It is hoped that multimodal and multimedia presentations of dictionary contents and the combination of several layers of information from different dictionary sources might improve the situation. In particular, an encoding or production dictionary that allows users to start from the source language and end up producing the target language correctly and accurately for the task at hand is still an unrealized goal. This is one of the least researched areas, and awaits further investigation and actual products.

SEE ALSO: Bilingual Lexicography; Corpus Linguistics: Overview; Technology and Terminology; Technology and Translation; Terminology and Data Encoding; Traditional Approaches to Monolingual Lexicography

References

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford, England: Oxford University Press.

CJK Dictionary Institute. (Comp.). (2009). *Japanese–English dictionary of technical terms*. Tohoku Niiza-shi, Japan: CJK Dictionary Institute.

Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research and Development*, *47*(3), 15–33.

Cobb, T. (2003). Do corpus-based electronic dictionaries replace concordancers? In B. Morrison, G. Green, & G. Motteram (Eds.), *Directions in CALL: Experience, experiments, evaluation* (pp. 179–206). Hong Kong: Polytechnic University.

Duval, A. (2008). Equivalence in bilingual dictionaries. In T. Fontenelle (Ed.), *Practical lexicography: A reader* (pp. 273–82). Oxford, England: Oxford University Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465–80.

Matsumoto, Y., & Utsuro, T. (2000). Lexical knowledge acquisition. In R. Dale, M. Hermann, & H. Somers (Eds.), *Handbook of natural language processing* (pp. 563–610). New York, NY: Marcel Dekker.

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston, MA: Heinle.

Rundell, M. (1998). Recent trends in English pedagogical lexicography. *International Journal of Lexicography*, 11(4), 315–42.

Svensén, B. (2009). *A handbook of lexicography*. Cambridge, England: Cambridge University Press.

Tono, Y. (2005). *Shogakukan corpus-based dictionary of English synonyms*. Tokyo, Japan: Shogakukan.

Tono, Y. (Ed.). (2008). *Ace Crown English–Japanese dictionary*. Tokyo, Japan: Sanseido.

Tono, Y. (2009). The potential of learner corpora for pedagogical lexicography. In V. Ooi, A. Pakir, I. Talib, & P. Tan (Eds.), *Perspectives in lexicography: Asia and beyond* (pp. 105–15). Jerusalem, Israel: K Dictionaries.

Suggested Readings

Fontenelle, T. (Ed.). (2008). *Practical lexicography: A reader*. Oxford, England: Oxford University Press.

Hartmann, R. R. K. (Ed.). (2003). *Lexicography: Critical concepts I–III*. London, England: Routledge.

Tono, Y. (2001). *Research on dictionary use in the context of foreign language learning*. Tübingen, Germany: Niemeyer.

Tono, Y. (2003). Learner corpora: Design, development and applications. In D. Archer, P. Rayson, A. Wilson, & A. McEnery (Eds.), *Proceedings of Corpus Linguistics 2003* (pp. 800–9). Lancaster, England: Lancaster University.