The Web as Corpus in Translation

Li-jue SONG
East China University of Political Science and Law, Shanghai, China
stella2002@163.com

Keywords: Web, Corpus, Translation.

Abstract. This thesis introduces corpus-based practices and studies in translation, including monocorpus, parallel corpus and web (as corpus) using in translation. Making an effective query on internet has been demonstrated in this thesis by WebCONC, Webcorp and KWICFinder alike, and 9 corpus tools have also been listed in the thesis to facilitate translation practices and studies.

Corpus Linguistics and Translation Studies

Almost 25 years have passed since Mona Baker’s (1993) seminal article on the application of insights from corpus linguistics to translation studies. Since then, corpus translational studies or corpus-based language instruction have become a standard component in many translation courses of Universities, and become a mainstream in descriptive translation studies. Currently, computational applications increasingly based on corpora such as translation memories (TMs) and machine translation (MT) systems have become part of life for all language services providers, not only for specialized and technical translators. Monographs and collected volumes have appeared on both theoretical and descriptive aspects (e.g. Laviosa 2002; Olonhan 2004; Anderman and Rogers 2008). Obviously, the creation of corpus along with the development of computer science has inevitably influenced the ecology of translation studies, theories and practices.

Generally, there are two types of studies by the way of corpora: the corpus-based study and the corpus-driven study. Corpus-based study mainly based on theories of linguistics and translation, the corpus used in the study can be regarded purely as a hybrid of theory and method. Nevertheless, the corpus-driven study is closely linked to the study of large data, which refers to large text-data in linguistics or translations. Especially, it is necessary to identify simplification, explicitation, and standardization by corpora. Apart from local corpora tools to index and calculate words and collocations, web has become a mega corpus with many sub-corpora which are needed to be mined with advanced tools access the Internet.

Web as Corpus

With different ways to access, display and interpret language data, web and corpora share more differences than common grounds. Furthermore, specified software is often used in corpora analysis, yet Web is accessed primarily through commercial search engines as Google, Bing or Baidu etc. The point is whether corpora software can be used for large data on net, or whether Web can be used for linguistic or translational research.

One of problems with Web searching is to make an effective query getting only relevant results. It is obvious that searching engines have been developed quickly with their new technical programming and simpler operations. Some general regular expressions are used in search like the way of corpus searching. Words and terms pertaining both to text type and content can be used in search to weed out most references to the keyword. Also there is a statistical result with the frequency of keywords in the front of every searching page. These quantitive results being expressed in number of hits only provide a rough if not confusing indication of standard usage, whereas the search engines fail to present linguistic analysis, in other way to say, either fail to present results with a format suitable for linguistics or translation.

The searching results consist in a list of links to the indexed documents containing the query
string, which lacks linguistic criteria for the commercial quality of the Web. That is to say, commercial considerations and other ranking criteria other than authoritativeness determined the finding results. Under this background, *WebCONC*, *Webcorp* (Huning 2001-2002) and *KWiCFinder* were designed and developed with the hope to utilize the big data from Web into linguistic studies. *Webcorp* acts as an interface between the end user and search engines, allowing the display of search results in a concordance-like format as in lines. It adds a module of ‘wordfilter’ to restrict the number of instances per site to one, and exclude hyperlink text, email addresses and other distracters from searches (Reounf *et al.* 2007). As well, *Webcorp* has a large-scale search engine with more search options, part-of-speech tags and quantitative analyses.

As to get the further manipulation of search results, a cache mode is used to store data on the server for a period of time and those search results can be downloaded. By doing which, the system basically creates a corpus from the Web from which concordance lines can be sorted, collocations generated and other relations between words can be computed (Zanettin 2011:60). *KWiCFinder* creates a cache copy of research results yet by a separated program which is needed to download and run for the desktop. With online tools we mentioned above, both the main information about wordlist, POS, collocation, and the overhead information about URL, time, textual domain can be collected as a source of corpora. From this perspective, the Web is not only as a corpus “surrogate” (Bernardini *et al.* 2006:10) but also as a source of texts for corpora. Since the data to the third-generation corpora has not been limited to the those printed but electronic texts, which is much more readily available than printed texts to save costs on labors and expenses of corpora construction.

**Corpus Annotation for Descriptive Translation Studies**

Basically, there are two types of annotation or markup, i.e. Procedural or presentational markup and descriptive or structural markup. Procedural markup concerns information about the visual formatting of a text, for instance regarding the font type and size of titles or the way in which “foreign words” are displayed in the text. Descriptive markup concerns instead the content of a document both as regards linguistic and extra linguistic information, such as bibliographic information about a text, information about its logical structure, and information about smaller textual units, down to the orthographic word and even to word parts (morphemes). As we all know, these linguistic information with different textual segments can be annotated by the way of automation or semi-automation, including tools for tokenization, POS tagging, lemmatization and syntactic annotation or parsing. When it comes to the level of discourse and pragmatic analysis, it may be added as further layers. Annotation can be applied to referring expressions (Biber *et al.* 1998:112-122) and for instance, cohesive devices, proper names and metaphorical expressions (Stefanowitsch and Gres 2006), even to the speech act categories and idiomatic expressions.

The annotated criteria and processing are different form monolingual corpus to translational corpus. On the structural level, it is needed to separate textual units, referring translators’ notes or a glossary whose annotation requiring additional considerations on the part of corpus compilers, from source texts to translated texts. Those structural and linguistic annotations may be exploited in translated corpora to draw alignment maps to be used in the retrieval of parallel concordances.

**XML**, eXtensible Markup Language, is often used as the structural form of annotation for distributed electronic texts. It is easy to learn and readable to computer with its simple code and strict syntax rules. It may look like the following:

```xml
<?xml version="1.0">
<doc>
  <title>This is the title</title>
  <p>This is the text</p>
</doc>
```

All elements are listed with the start marker “<*>” and the end marker “</*>”, and herein “*” means any xml syntactic markers, like <p> for paragraph, <title> for headline. Each element can be expanded to contain further parameters, called attributes, which specify properties of the element.
XML documents are expressed in programming languages such as DTD (Document Type Definition) language being settled under the TEI annotation framework, and are analyzed by the searching method of Web browser or concordance. At the same time, XML provides the syntax for annotation information about texts in a way of automatic analyzing by corpus management software. It is needed to note a TEI header element `<teiHeader>` may contain four main elements: “file description `<fileDesc>”’, “encoding description `<encodingDesc>`”, “profile description `<profileDesc>`” and “revision description `<revisionDesc>`”. Those annotations give much information on texts, like the `<sourceDesc>` element contains not only information about the translated novel but also bibliographic information about the original language text. This is contained in the `<relatedItem>` element nested within the main bibliographic `<bibl>` element.

Translational texts are more complicated than mono-lingual texts, a classification declaration `<classDecl>` element referring to and assigning descriptive labels to the classificatory categories used for the texts in the corpus, or translation status are needed.

With annotating source language and target language(s) respectively on different layers, different documentary information, structural information and text-linguistic information within two languages will be identified by concordancer or XAIRA. Texts of different genres are differ much from one to another, annotations helps us to excavate distinct syntactic, lexical, or semantic features of them and their translated texts. The premises are those tools and analysis of data.

**Corpus Analysis with Different Tools**

A stable corpus used in translation studies, practices and pedagogical working is workable and efficient. To examines the main types of computer-aided analyses and translation practice which can be carried out using corpus tools. It is needed to differ monolingual corpora from parallel and comparable corpora using in translation studies, practices and teaching.

Text acquisition is the first stage in corpus compilation with the way of obtaining clean, text-only versions of the documents to be included in a corpus. The process can be done either by digitalizing hard works or converting existing electronic documents from different format (PDF, HTML, etc). Different software applications will be used in this stage, depending on different formats. Meanwhile, contextual and metatextual information about each document can be entered through an input template and stored in a standard header, associated with a text which can be marked up though a mix of manual and automatic means. Freely available XML editors can be found by searching sites such as SourceForge (http://sourceforge.net/) and CNET.com (http://download.cnet.com). XML copy editor, Rinzo XML editor, XML Marker, and XML Pad, to mention only a couple. A popular free application is the Tree Tagger which is a POS tagger and a lemmatizer, adaptable to different languages on the basis of a lexicon and a manually tagged training corpus, and it has been used to tag both English and Chinese texts. Besides, XAIRA and MODNLP-tec include client applications which are installed on local workstations and used for corpus interrogation and data display, and other applications do not provide a graphical user interface.

Other than data retrieval and annotation, some basic statistics of wordlists, concordances, collocations and colligations are also need to be considered as the essential result of translational and linguistic studies. Word types and tokens are two basic factors for textual analysis. Thusly the ratio of types and tokens (TTR) is used to evaluate the difficult of source texts or target texts (e.g. TTR 40% in source language yet 60% in target language will demonstrate the complex strategy using in the process of translation). A comparison between wordlists from different texts or corpora can also provide quite interesting data. Two or pairs of different translations of the same source text(s) can be compared to extract a list of words that are ‘key’ to each translated text (or corpus). These are the words whose frequency in one text is significantly higher than that of the same words in the other texts. Many tools are used to generate word lists, concordances, collocations and colligations, like Antconc (a monolingual software for free), Paraconc (a multilingual business software), Wordsmith (a monolingual business software) and many other online applications. Those
terms in corpus studies means much in analyzing texts and is very helpful in translation studies. Comparable and parallel corpora are no doubly play an important role in descriptive studies of translation, as they represent the backdrop against which features of source and target texts in parallel corpora can be accessed.

Similar or equivalent features across languages in both comparable and parallel corpora may also be investigated by looking at concordances in each subcorpus independently (e.g. Kenny 2000). To make translation studies become more detailed and translation practices become more effective, all software can be used to build corpora listed as follows:

1. Alinea, Windows application for interactive sentence alignment, language independent but includes optional parameters for specific language pairs, freeware, at http://w3.u-grenoble3.fr/kraif/.
3. InterText, an editor for aligned parallel texts which build on the TCA2 and hunalign automatic aligners (developed by Pavel Vondřič).
4. NATools, a workbench for parallel corpora processing which includes a sentence aligner and a Probilistic Translation Dictionary extractor, a word aligner and a set of other tools, Unix environment, open source software, at http://sourceforge.net/projects/natools/.
9. Uplug, a suite of tools for pre-processing parallel corpora. Contains a tokenizer, a sentence-splitter, XML-tools, a sentence aligner, a word aligner, a corpus indexer (using the CWB), and Web search interfaces. Open source, at http://stp.lingfil.uu.se/joerg.uplug/.

Conclusion

As the main method of descriptive translational studies, comparable corpora can be used by translation trainees and professionals include large as well as smaller specialized corpora. Large corpora mean large data describing the true language by various corpora tools. It is necessary to know different software and apply them in translation. This paper introduces and integrates different software of corpora (the monolanguage and parallel languages) to facilitate translation practice or studies. Different stages and tools involved in corpus compilation and use are outlined, from corpus encoding and annotation to indexing and data retrieval, and various methods and techniques that allow end users to make sense of corpus data are described.

References


241


