Harvesting the Bitexts of the Laws of Hong Kong From the Web

Chunyu Kit Xiaoyue Liu KingKui Sin Jonathan J. Webster
Department of Chinese, Translation and Linguistics
City University of Hong Kong, Tat Chee Ave., Kowloon, Hong Kong
{ctckit, xyliu0, ctsinkk, ctjjw}@cityu.edu.hk

Abstract

In this paper we present our recent work on harvesting English-Chinese bitexts of the laws of Hong Kong from the Web and aligning them to the subparagraph level via utilizing the numbering system in the legal text hierarchy. Basic methodology and practical techniques are reported in detail. The resultant bilingual corpus, 10.4M English words and 18.3M Chinese characters, is an authoritative and comprehensive text collection covering the specific and special domain of HK laws. It is particularly valuable to empirical MT research. This piece of work has also laid a foundation for exploring and harvesting English-Chinese bitexts in a larger volume from the Web.

1 Introduction

Bitexts, also referred to as *parallel texts* or *bilingual corpora*, collections of bilingual text pairs aligned at various levels of granularity, have been playing a critical role in the current development of machine translation technology. It is such large data sets that give rise to the plausibility of empirical approaches to machine translation, most of which involve the application of a variety of machine learning techniques to infer various types of translation knowledge from bitext data to facilitate automatic translation and enhance translation quality. Large volumes of training data of this kind are indispensable for constructing statistical translation models (Brown et al., 1993; Melamed, 2000), acquiring bilingual lexicon (Gale and Church, 1991; Melamed, 1997), and building example-based machine translation (EBMT) systems (Nagao, 1984; Carl and Way, 2003; Way and Gough, 2003). They also provide a basis for inferring lexical connection between vocabularies in cross-languages information retrieval (Davis and Dunning, 1995).

Existing parallel corpora have illustrated their particular value in empirical NLP research, e.g., Canadian Hansard Corpus (Gale and Church, 1991b), HK Hansard (Wu, 1994), *INTERSECT* (Salkie, 1995), *ENPC* (Ebeling, 1998), the Bible parallel corpus (Resnik et al., 1999) and many others. The Web is being explored not only as a super corpus for NLP and linguistic research (Kilgarriff and Grefenstette, 2003) but also, more importantly to MT research, as a treasure for mining bitexts of various language pairs (Resnik, 1999; Chen and Nie, 2000; Nie and Cai, 2001; Nie and Chen, 2002; Resnik and Smith, 2003; Way and Gough, 2003). The Web has been the playground for many NLPers. More and more Web sites are found to have cloned their Web pages in several languages, aiming at conveying information to audience in different languages. This gives rise to a huge volume of wonderful bilingual or multi-lingual resources freely available from the Web for research. What we need to do is to harvest the right resources for the right applications.

In this paper we present our recent work on harvesting English-Chinese parallel texts of the laws of Hong Kong from the Web and construct-

ing a subparagraph-aligned bilingual corpus of about 20 million words. The bilingual texts of the laws is introduced in Section 2, with an emphasis on HK's legislation text hierarchy and its numbering system that can be utilized for text alignment to subparagraph level. Section 3 presents basic methodology and technical details for harvesting and aligning bilingual Web page pairs, extracting content texts from the pages, and aligning text structures in terms of the text hierarchy via utilizing consistent intrinsic features in the Web pages and content texts. Section 4 presents XML schema for encoding the alignment results and illustrates the display mode for browsing the aligned bilingual corpus. Section 5 concludes the paper, highlighting the value of the corpus in term of its volume, translation quality, specificity and comprehensiveness, and alignment granularity. Our future work to explore the Web for harvesting more quantities of parallel bitexts is also briefly outlined.

2 Bilingual Texts of the Laws of HK

The laws of Hong Kong (HK) before 1987 were exclusively enacted in English. They were translated into Chinese in the run-up to the handover in 1997. Since then all HK laws have been enacted in both English and Chinese, both versions being equally authentic. This gives rise to a valuable set of bitexts in large quantity and high quality that can be utilized to facilitate empirical MT research.

2.1 BLIS Corpus

The bilingual texts of the laws of Hong Kong have been made available to the public in recent years by the Justice Department of the HKSAR through the bilingual laws information system (BLIS). All these texts are freely accessible from `http://www.justice.gov.hk/`.

BLIS provides the most comprehensive documentation of HK legislation. It contains all statute laws of Hong Kong currently in operation, including all ordinances and subsidiary legislation of HK (and some of their past versions dating back to 60 June 1997), the Basic Law and the Sino-British Joint Declaration, the constitution of PRC and national laws that apply in HK, and other relevant instruments. The entire bilingual corpus of

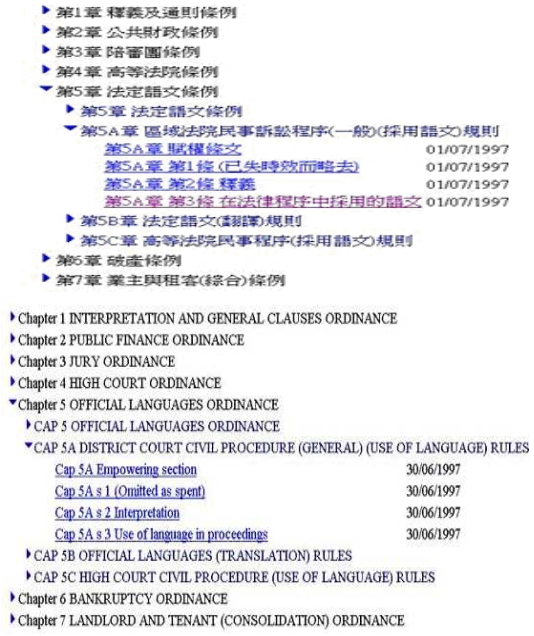

Figure 1: Illustration of BLIS hierarchy

BLIS legal texts contains approximately 10 million English words and 18 million Chinese characters. Lexical resources of this kind are particularly useful in bilingual legal terminology studies and text alignment work.

2.2 Text Hierarchy

BLIS organizes the legal texts in terms of the hierarchy of the Loose-Leaf Edition of the Laws of Hong Kong. At the top level, the ordinances are arranged by chapters, each of which is identified by an *assigned number* and a *short title*, e.g., `Chapter 5 OFFICIAL LANGUAGES ORDINANCE / 第5章 法定語文條例`. The assigned number for a subsidiary legislation chapter consists of a chapter number and a following uppercase letter, e.g., `CAP 5C HIGH COURT CIVIL PROCEDURE (USE OF LANGUAGE) RULES / 第5C章 高等法院民事程序(採用語文)規則`.

The content of an ordinance, exclusive of its long title, is divided and identified according to a very rigid numbering system which encodes the hierarchy of the texts of the laws. Both the Chinese and English versions of an ordinance follow exactly the same hierarchical structures such as chapters (章), parts (部), sections (條), subsections (款), paragraphs (段) and subparagraphs (節). This allows us to align the bitexts along

章：	5A	標題：	區域法院民事訴訟程序(一般)(採用語文)規則	憲報編號：	25 of 1998 s. 2
條：	3	條文標題：	**在法律程序中採用的語文**	版本日期：	01/07/1997

附註：
具追溯力的修訂—見1998年第25號第2條

(1) 法官可爲在其席前進行的任何法律程序獲公正而迅速地處理，而在該法律程序或其任何部分中，按他認爲適當兼用兩種法定語文或採用其中一種。
(2) 法官根據第(1)款作出的決定是最終決定。
(3) 在法院進行的法律程序中或法律程序的一部分中的一方，或在法院進行的法律程序中或法律程序的一部分中的證人—

(a) 可兼用兩種法定語文或採用其中一種；及
(b) 可用任何語文向法院陳詞或作供。

(4) 在法院進行的法律程序中或法律程序的一部分中的法律代表，可兼用兩種法定語文或採用其中一種。
(5) 供法院在任何法律程序中使用的文件，可用其中一種法定語文製備。

Chapter:	5A	Title:	DISTRICT COURT CIVIL PROCEDURE (GENERAL) (USE OF LANGUAGE) RULES	Gazette Number:	
Section:	3	Heading:	**Use of language in proceedings**	Version Date:	30/06/1997

(1) A judge may use either or both of the official languages in any proceedings or a part of any proceedings before him as he considers appropriate for the just and expeditious disposal of the proceedings before him.
(2) The decision of the judge under subrule (1) is final.
(3) A party to or a witness in any proceedings or a part of any proceedings before the court may-

(a) use either or both of the official languages; and
(b) address the court or testify in any language.

(4) A legal representative in any proceedings or a part of any proceedings before the court may use either or both of the official languages.
(5) Documents prepared for use by the court in any proceedings may be in either official language.

Figure 2: BLIS texts in pair

this hierarchical structure, once they are downloaded from the BLIS official site. To our knowledge, a well-aligned bilingual corpus of this size covering a special domain so comprehensively is seldom readily available for the Chinese-English language pair.

Excerpts from the BLIS corpus are illustrated in Figure 1 and 2, one illustrating its hierarchy and the other a pair of BLIS bitexts. From the excerpts we can see that not everything has an exact match between a pair of BLIS Web pages. For example, the Chinese side has a gazette number "25 of 1998 s. 2" and a piece of "remarks" at the beginning of content text, whereas its English counterpart has none of them.

3 Harvesting Bitexts from the Web

Basically two phases are involved in constructing the bilingual corpus of the laws of HK. The first phase is to harvest the monolingual texts of HK laws from the BLIS site and align them into pairs. It involves the following steps: (1) downloading Web pages one by one with the aid of a Web crawler, (2) extracting the texts from them by filtering out the HTML markup, and (3) aligning the extracted monolingual texts into bilingual

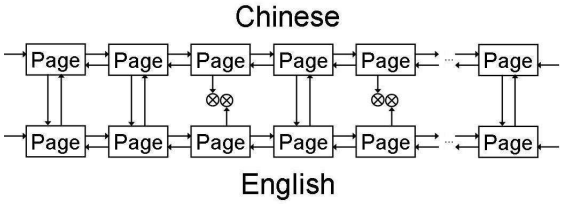

Figure 3: BLIS web pages connected as two double linked lists

pairs. The second phase is to align finer-grained text structures within each text pair.

3.1 Downloading BLIS Web Pages

A BLIS Web page does not necessarily correspond to any particular text structure such as a chapter, a part, a section, a subsection, or a paragraph in the BLIS hierarchy. A chapter, especially a short one, may be organized into a few sections in a Web page or in several contiguous pages. Some sections, e.g., the long ones, are divided into several pages. In general, BLIS does not maintain any reliable match between its Web pages and any particular text hierarchical structures.

Fortunately, in most cases a BLIS page always has a counterpart in the other language. There is a "switch language" button on each page to link to the counterpart page. Such linkage allows us to download the Web pages in pairs and, consequently, harvest a list of page-to-page aligned bitexts.

In addition to the pair link, each BLIS page also carries links for the "next" and the "previous section of enactment". These two kinds of linkage turn the pages into two double linked lists, each in a language, as illustrated in Figure 3, with each page as a node. Nodes in pairs are also double linked between the two lists.

However, the pairwise linkage is not reliable in the BLIS site, because there are missing Web pages in one of the two languages in question (see Table 3 below for more details). In order to download all bitexts of legislation from the site, we need to go through one linked list and download each page and its counterpart, if there is one, in the other language. Such scanning gives a list of text pairs, where some pages may have a null

	Total time	Downloaded files
English	17 hours	50,638 (429MB)
Chinese	18 hours	50,510 (460MB)

Table 1: File downloading

	File name	
BLIS HTML page title	Chinese	English
`Cap 5A ...`	`5A_c.txt`	`5A_e.txt`
`Cap 5A s 1 ...`	`5A-1_c.txt`	`5A-1_e.txt`
`Cap 5A s 2 ...`	`5A-2_c.txt`	`5A-2_e.txt`
`Cap 5A s 3 ...`	`5A-3_c.txt`	`5A-3_e.txt`

Table 2: Naming downloaded files in terms of BLIS numbering

counterpart. An alternative strategy is to download each list separately, and then match the pages into pairs sequentially with the aid of numbering information in the header of each page – see 3.2 below. These two strategies verify one another, making sure that all pages are downloaded and put in the right pairs.

The downloading is carried out by a Web crawler implemented in Java. In order to accomplish the above strategies, it also has to handle a number of technical issues.

- It sleeps for a while (e.g., 10 seconds) when it finishes downloading a certain number of pages (e.g., 50 pages), because the BLIS site refuses continuous access from one site for a too long time.
- When an error occurs, it remembers the current URL. Then it re-starts from where it stops.

The data about the file downloading from BLIS site is given in Table 1. One can conceive that if the time intervals for sleep and downloading could be automatically tuned by the crawler to maximize the downloading efficiency, it would get the job done significantly more quickly. Our option for 10 seconds sleep between every 50 files is based on error records of a number of test runs.

3.2 Aligning Web Pages

Every BLIS Web page is identified by a subtitle that carries numbering information about the page, as illustrated in Figure 1. Such a subtitle is exactly retained in the page as its HTML title.

Files	English	Chinese
Aligned	50,506 (62.3MB)[a]	50,506 (38.5MB)
Missing	132	4
Total	50,638	50,510
Size[b]	10.4M words	18.3M char.s

[a]The size of extracted texts.
[b]Exclusive of punctuation marks.

Table 3: The number of aligned and missing files

This feature is utilized to align BLIS pages: all downloaded files are named in terms of the numbering information extracted from their HTML titles, as illustrated in Table 2. Consequently, all files are naturally aligned in pairs by their names. Any file names not in a pair indicate the missing counterparts in the other language. The statistics of file alignment are given in Table 3.

3.3 Text Extraction

Basically, this task involves two aspects, namely, filtering HTML markup and extracting content text. A straightforward strategy is that we first clean up HTML tags in each page and then the non-legal content. The tags are in brackets, and non-legal content in a consistent pattern throughout all BLIS pages. However, a more convenient way to do it is to make use of a reliable feature in the BLIS pages: legal content is placed in between two – the only two – horizontal bars in each page. Accordingly, we implement a strategy to first extract every thing in between the two bars and then clean up remaining HTML tags. The output from this procedure includes

- a header as a fixed set of items, including chapter number, title, heading, etc., and
- a piece of content text as a list of numbered items each in a line. (See the header and content text in Figure 2.)

The text in a BLIS page is displayed as a sequence of hierarchically numbered items, such as subsections, paragraphs and subparagraphs.

3.4 Text Alignment within Text Pairs

After page (or file) alignment, each page finds its counterpart in the other language. After text extraction, a page gives a content text consisting of a list of numbered items, each in a line. A such

```
Remarks:
Adaptation amendments retroactively made - see
26 of 1999 s.3//[a]
(1) All Ordinances shall be enacted and published
    in both official languages.//
(2) Nothing in subsection (1) shall require an
    Ordinance to be enacted and published in
    both official languages where that Ordinance
    amends another Ordinance and-//
(a) that other Ordinance was enacted in the
    English language only; and//
(b) no authentic text of that Ordinance has been
    published in the Chinese language under
    section 4B(1).//
(3) Nothing in subsection (1) shall require an
    Ordinance to be enacted and published in both
    official languages where the Chief Executive
    in Council- (Amended 26 of 1999 s.3)//
```

[a] Indicating a text line break.

Table 4: Anchors in a sample text

item can be divided into a numbering item and the remaining content text in the line, as illustrated in Table 4. The Chinese counterpart of this text carries similar lines, if no missing line in any page of the pair.

Unfortunately, missing lines are found in some BLIS pages, as exemplified in Figure 2. There is no guarantee that matching text lines one by one in sequence would carry out the expected alignment within a page pair. However, the numbering items at the beginning of each line can be utilized as *anchors* to facilitate the alignment. The strategy along this line is given as follows.

1. Anchor identification: numbering items *at the beginning of each line* are recognized as anchors, with the beginning and the end of the whole content text as two special anchors, resulting in a list of anchors for each page;

2. Anchor alignment: match the two lists of anchors sequentially. If a pair of anchors does not match, give up the smaller one (in terms of the BLIS numbering hierarchy) and move on to the next possible pair, working in exactly the same procedure as matching identical anchor pairs between two sorted lists of anchors.

3. Text line alignment: a pair of matched anchors give a pair of matched lines; an unmatched anchor indicates a missing line in the other language.

4 XML Markup for the Aligned Corpus

XML is applied to encode the text alignment outcomes output from the above alignment procedure. It has been a standard for data representation and exchange on the Web, and also accepted by the NLP community as a standard for linguistic data annotation and representation (Ide et al., 2000; Mengel and Lezius, 2000; Kim et al., 2001). There are a series of yearly NLPXML workshops for it since 2001. It provides a platform-independent flexible and sophisticated plain text format for data encoding and manipulation. It is particularly suitable for hierarchical linguistic data such as the hierarchically-aligned bilingual corpus that we have produced. What's more, converting data to XML format not only significantly reduces the complexity of data exchange among different computer systems but also enhances data transmission reliability and eases Web browsing.

There have been many corpora that are annotated with XML, e.g., HCRC Map Task Corpus (Anderson et al., 1991), American National Corpus (Ide and Macleod, 2001), the La Republica corpus (Baroni et al., 2004). Below we present the XML schema for our subparagraph-aligned BLIS bitexts, with sample annotation, and necessary Web browsing.

4.1 XML Schema

The current version of the XML schema for the bilingual BLIS corpus, as given in Figure 4, focuses on encoding all text structures in the BLIS hierarchy, including all elements in each BLIS Web page. It is to be extended to cover finer-grained structures such as clauses, phrases and words, as we proceed to align the BLIS bitexts at these linguistic levels. For simplicity, we allow `para` to subsume all types of text line, be they a section, subsection, paragraph or subparagraph. The annotation of a sample bitext with this schema is illustrated in Figure 5. Annotation of this kind is carried out by a Java program automatically for the entire bitext corpus.

4.2 Corpus Browsing

A number of display modes are designed for browsing the subparagraph-aligned bitexts, including bilingual modes and monolingual modes.

Figure 4: XML schema for aligned BLIS bitexts

In a bilingual mode, text line pairs are displayed in sequence. Switch of language order or from one mode to another is allowed any time during browsing. The bilingual display mode is illustrated in Figure 6.

5 Conclusion

We have presented in the above sections our recent work on harvesting and aligning the bitexts of the laws of Hong Kong, including basic techniques for downloading English-Chinese bilingual legal texts from BLIS official site, sound strategies for aligning the bitexts by utilizing the numbering system in the legal texts, and necessary XML annotation for the alignment results. The value of the outcomes, i.e., the subparagraph-aligned bilingual corpus, can be evaluated in terms of the following aspects.

Corpus size The entire corpus is of 10.4M English words and 18.3M Chinese characters, several times larger than the well-known Penn Treebank Corpus in size.

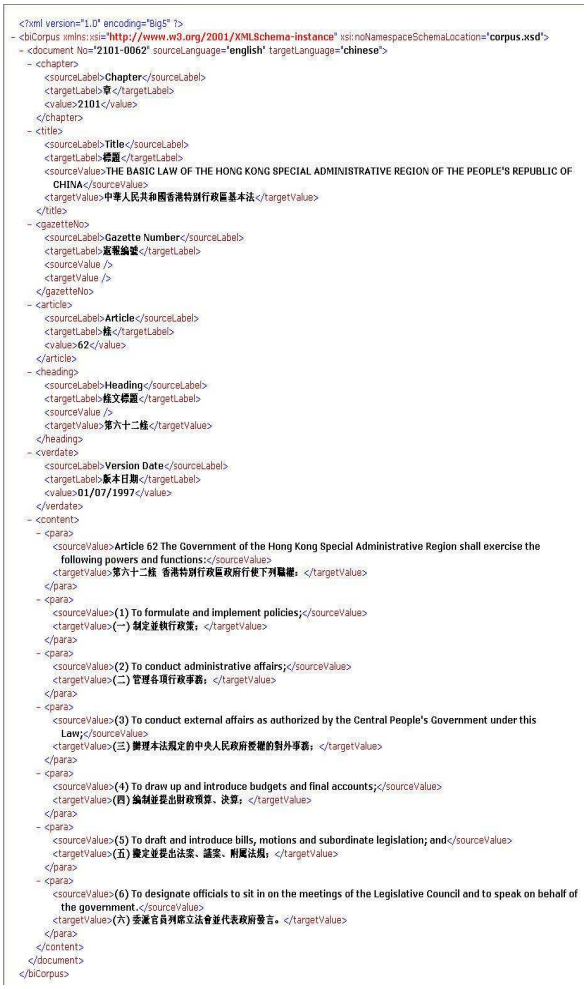

Figure 5: Sample bitext in XML encoding

Translation quality All texts of the corpus are prepared by the Law Drafting Division of the Department of Justice, Hong Kong Government. Legal texts are known to be more precise and less ambiguous than most other types of text.

Specificity and comprehensiveness The corpus covers specifically the domain of Hong Kong legislation. It is the most authoritative and complete text collection of the laws of Hong Kong.

Alignment granularity The entire corpus is aligned precisely to the subparagraph level. Most subparagraphs in the legal texts are phrases, fragments of a clause, or clauses; as shown in Table 4.

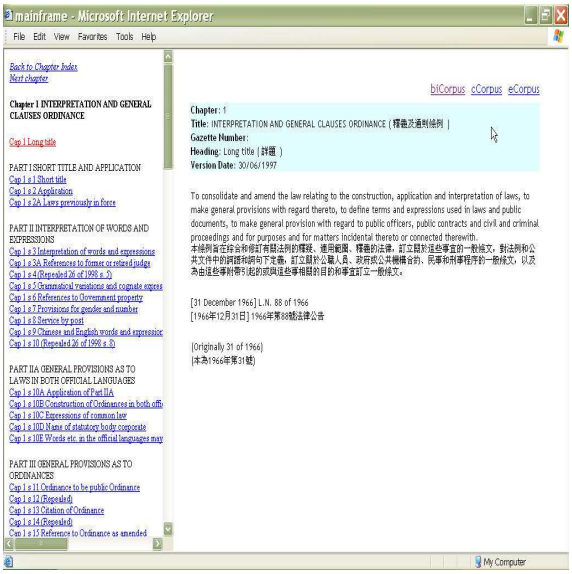

Figure 6: Illustration of browsing modes

A bilingual corpus of this size and quality covering a specific domain so comprehensively is particularly useful not only in empirical MT research but also in computational studies of bilingual terminology and legislation. Our future work will focus on word alignment for inferring bilingual lexical resources and on automatic recognition of legal terminology.

Also, our experience in constructing this bilingual corpus has laid a foundation for us to continue to harvest more bilingual text materials from the Web, e.g., from Hong Kong government's Web sites. We find that almost all Hong Kong government web sites, which are in large numbers, maintain their Web pages consistently parallel in English and Chinese. We are not sure if such bitexts in such pages are larger than that in the BLIS site in volume. We do know they cover a large number of distinct domains. This is particularly useful for MT. If we can harvest and align the bitexts from such Web pages efficiently via utilizing their intrinsic characteristics of URL correspondence and text structure, it would not be a dream any more to put an end to the time of having too few existing translation materials for empirical MT studies, at least, for the language pair of Chinese and English.

Acknowledgements

The work described in this paper was supported by the Research Grants Council of HKSAR, China, through the CERG grants 9040861 and 9040482. We wish to thank our team members for their help.

References

Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In *LREC 2004*, pp. 1771-1774.

Simon P. Botley, Anthony M. McEnery, and Andrew Wilson (eds.). 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Michael Carl and Andy Way (eds.). 2003. *Recent Advances in Example-based Machine Translation*. Dordrecht: Kluwer.

Jiang Chen and Jian Y. Nie. 2000. Parallel Web text mining for cross-language information retrieval. In *RIAO'2000*, pp. 62–77. Paris.

Mark Davis and Ted Dunning. 1995. A TREC evaluation of query translation methods for multi-lingual text retrieval. In *TREC-4*, pp. 483–498. NIST.

Jarle Ebeling. 1998. Contrastive linguistics, translation, and parallel corpora. In *Meta*, 43(4):602–615.

William A. Gale and Kenneth W. Church. 1991. Identifying word correspondences in parallel texts. In *Fourth DARPA Workshop on Speech and Natural Language*, pp. 152–157. Asilomar, California.

William A. Gale and Kenneth W. Church. 1991b. A Program for Aligning Sentences in Bilingual Corpora. In *ACL'91*, pp. 177–184. Berkeley.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: an XML-based encoding standard for linguistic corpora. In *LREC2000*, pp. 825–830. Athens, Greece.

Nancy Ide and Catherine Macleod. 2001. The American National Corpus: A Standardized Resource of American English. *Proceedings of Corpus Linguistics 2001*, Lancaster UK.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. 2001. XML-based linguistic annotation of corpus. In *NLPXML-1*, pp. 47–54. Tokyo.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *EMNLP'97*, pp. 97–108. Brown University, August.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Andreas Mengel and Wolfgang Lezius. 2000. An XML-based representation format for syntactically annotated corpora. In *LREC2000*, Volume 1, pp. 121–126. Athens, Greece.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and Human Intelligence*, pp. 173–180. Amsterdam: North-Holland.

Jian Y. Nie and Jian Cai. 2001. Filtering noisy parallel corpora of Web pages. In *IEEE Symposium on Natural Language Processing and Knowledge Engineering*, pp. 453–458. Tucson, AZ.

Jian Y. Nie and Jiang Chen. 2002. Exploiting the Web as Parallel Corpora for Cross-Language Information Retrieval. *Web Intelligence*, pp. 218–239.

Philip Resnik, Mari B. Olse, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the "Book of 2000 Tongues". *Computers and the Humanities*, 33(1-2):129–153.

Philip Resnik. 1999b. Mining the Web for Bilingual Text. In *ACL'99*, pp. 527–534. Maryland.

Philip Resnik and Noah A. Smith. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

Raphael Salkie. 1995. INTERSECT: a parallel corpus project at Brighton University. *Computers and Texts 9* (May 1995), pp. 4–5.

Jean Veronis. 2000. *Parallel Text Processing*. Dordrecht: Kluwer.

Andy Way and Nano Gough. 2003. wEBMT: Developing and validating an example-based machine translation system using the World Wide Web. *Computational Linguistics*, 29(3):421–457.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *ACL'94*, pp. 80–87. Las Cruces, New Mexico, U.S.A.