# Developing Asian language corpora: standards and practice

**Zhonghua Xiao, Tony McEnery, Paul Baker, Andrew Hardie**
Department of Linguistics
Lancaster University
Lancaster
{z.xiao, a.mcenery, j.p.baker, a.hardie}
@lancaster.ac.uk

## Abstract

This paper first discusses standards for developing Asian language corpora so as to facilitate international data exchange. Following this, we present two corpora of Asian languages developed at Lancaster University – the EMILLE Corpus, which contains 14 South Asian languages, and the Lancaster Corpus of Mandarin Chinese. Finally, we will demonstrate how to explore these corpora using *Xara* and other corpus tools.

## 1 Introduction

In the recent past, the focus of corpus linguistics was on the creation and exploitation of English language corpora. While many European language corpora have also been created and made available recently,[1] less progress has been made in the creation of Asian language resources. In contrast to the wide variety of languages in Asia, the number of publicly available Asian language corpora is very limited. This situation needs to be improved.

There are some Asian language corpora scattered around the world. Of these, Chinese is probably the language for which most corpus data is available. The Institute of Computational Linguistics of Peking University released a corpus containing one million words using one month's news texts from *People's Daily* (January 1998).[2] The PH corpus, compiled by Guo Jin, contains around two million words of newswire texts from the Xinhua News Agency (1990 – 1991).[3] Academia Sinica also released a five million word balanced corpus of Mandarin Chinese as used in Taiwan.[4] The LIVAC synchronous corpus of Chinese, created by City University of Hong Kong, is near completion.[5] A spoken Chinese corpus of situated discourse is under construction under the auspices of the Chinese Academy of Social Science (see Gu 2002). The LDC has also released some corpora of news texts and telephone conversations in Chinese.

Corpora for other Asian languages are relatively few, though there are some resources, notably for East Asia and Thailand, e.g. Korean (cf. Rim 2001), Japanese (cf. Shirai 2001; Goto et al 2001) and Thai (cf. Sornlertlamcanich 2001; Thongprasirt et al 2001). With a few exceptions, e.g. the Sinica Corpus for Chinese (see Huang 2001) and the Sejong Corpus for Korean, most Asian language corpora are specialized corpora (e.g. newspaper corpora).

The problems facing Asian corpus linguistics are self-evident, as are their causes. First of all, there appears to be a lack of coordination in the development of Asian language resources. There is no true parallel to the European Language Resources Association (ELRA) in Asia, which could coordinate worldwide Asian language resource development efforts. There is also a need to establish standards and guidelines for corpus encoding

[1] See http://www.eida.fr/cata/tabtxt1.html, the website of Evaluations and Language Resource Distribution Agency.
[2] See http://icl.pku.edu.cn/Introduction/corpustagging.htm.
[3] A brief description of the corpus can be accessed online at ftp://ftp.cogsci.ed.ac.uk/pub/chinese/.
[4] See http://www.sinica.edu.tw/SinicaCorpus. The Sinica Treebank is accessible at http://140.109.19.103/treesearch.
[5] See http://www.livac.org/ for details.

and exchange in the region.[6] These standards, while taking into account the unique features of Asian languages, must also conform to other major standards (e.g. the ELRA standards for corpus validation) so as to facilitate the exchange of corpus resources in the region.

This paper first seeks to propose standards regarding corpus structure, markup and character encoding for Asian languages. Following this, two corpora are presented, namely, the EMILLE Corpus, which contains 14 South Asian languages, and a balanced corpus of written Chinese, the Lancaster Corpus of Mandarin Chinese. We will then demonstrate how to explore these corpora using *Xara* (XML-aware *SARA*) and other corpus tools.

## 2 Standards of corpus development

This section discusses standards of corpus development. As noted in section 1, while there are some corpora available for Asian languages, standards for corpus constituents, data formats, file structure, markup, annotation and character encoding are clearly needed for efficient data exchange internationally. These are currently lacking, or not applied if they exist.

*Corpus constituents and data formats.* A corpus can consist of written text, transcribed speech or multimedia like audio/video clips; it can also be distributed using various media (e.g. disks, CD, DVD, tape) or online. Whatever form or medium it takes, it is only considered complete when all of the necessary constituents are available. They include primary (corpus files) and ancillary (documentation) components. The following data formats are commonly used, and are recommended for most corpus interchange tasks: XML/SGML for text files, MP3/WAV for audio files, MPEG/Quicktime for video files and PNG/JPG for image files. For documentation, open formats such as PDF, HTML, or XML are recommended because non-standard or proprietary formats which may require the use of specific software should be avoided as far as possible.[7]

*File structure, markup and annotation.* Each corpus file should consist of two parts: header and body. The header part provides metadata about the corpus file while the body part contains the corpus data proper. As an international standard, XML has proved to be a sound basis for standardizing corpus and annotation formats to facilitate easy data linkage and transformation (cf. Ide 2000: 2).

As for the header part, a number of metadata sets have been proposed in Europe and the US by, for example, the Dublin Core Metadata Initiative (DC), the Open Language Archives Community (OLAC), the ISLE Metadata Initiative (IMDI), MPEG7 and EAGLES (Corpus Encoding Standard or CES). DC provides 15 elements used primarily to describe authored web resources. OLAC is an extension of DC, which introduces refinements to narrow down the semantic scope of the DC elements and adds an extra element to describe the language(s) covered by the resource (cf. Wittenburg et al 2002: 1321). MPEG7 is principally oriented towards multimedia rather than textual data. Hence, many of its elements are not relevant to text corpora. While IMDI applies to (multimedia) corpora and lexica as well, it nevertheless needs special software (e.g. the IMDI BCEditor) to work efficiently. The standard we have used in our Asian corpus building work is the Corpus Encoding Standard (CES) developed by EAGLES.

The CES is an application of SGML (ISO 8879) compliant with the specifications of the TEI Guidelines for Electronic Text Encoding and Interchange of the Text Encoding Initiative. It is increasingly recognized as a standard for corpus building, with projects such as MULTEXT, PAROLE, BAF, TALANA and the American National Corpus project adhering to it (cf. Baker et al 2002). The CES specifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation as well as general architecture. Three levels of text standardization are specified in the CES: 1) the metalanguage level, 2) the syntactic level and 3) the semantic level. Standardization at the metalanguage level regulates the form of the syntactic rules and the basic mechanisms of markup schemes. Users can use a TEI-compliant Document Type Definition (DTD) to define tag names as well as "document models" which specify the relations among tags. As texts may still have different document structures and markups even with

[6] The ALR committee is working in the right direction by setting up a language resource repository and related standards.

[7] The discussion in this paragraph is based on ELRA document D1: *Validation Manual for Written Language Resources* (http://www.oucs.ox.ac.uk/rts/elra/D1.xml).

the same metalanguage specifications, standardization at the syntactic level specifies precise tag names and syntactic rules for using the tags as well as constraints on content. However, even the same tag names can be interpreted differently by the data sender and receiver. This is why standardization at the semantic level is useful. The CES seeks to standardize at the semantic level for those elements most relevant to language engineering applications, in particular, linguistic elements. The three levels of standardization are designed to achieve the goal of universal document interchange. In addition, the CES also provides encoding specifications for linguistic annotation (e.g. paragraph and sentence boundaries and morphological information for tokens) together with a data architecture for linguistic corpora. [8]

*Character encoding.* Character encoding is another area that needs to be standardized for corpus construction. In many cases, multiple and often competing encoding systems complicate Asian corpus building, providing a real problem as McEnery and Xiao (2004) observe. The main difficulty in building a multilingual corpus of Asian languages is the need to standardize the language data into a single character set (see McEnery et al 2001). We recommend Unicode as a solution to this problem. Unicode is truly multilingual in that it can display characters from a very large number of writing systems. From the Unicode Standard version 1.1 onwards, Unicode is fully compatible with ISO 10646-1 (UCS). The combination of Unicode and XML is a general trend in corpus development. As such it is to be welcomed.

The EMILLE and LCMC corpora were developed at Lancaster University following these guidelines. These corpora will be presented in the following two sections as an example of what can be achieved by adherence to these standards.

## 3 The EMILLE Corpus

The EMILLE Corpus is the primary resource developed by the EMILLE (Enabling Minority Language Engineering) project. [9] The corpus consists of three components: monolingual, parallel and annotated corpora. There are 14 monolingual corpora, including both written and (for some languages) spoken data for 14 South Asian languages. The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. The annotated component includes the Urdu monolingual and parallel corpora annotated for parts-of-speech, together with 20 written Hindi corpus files annotated to show the nature of demonstrative use. The rationale for the corpus design was explained and justified in McEnery et al (2001) and Baker et al (2002, 2003).

Table 1: The EMILLE monolingual corpora

| Language | Written | Spoken | Total |
|---|---|---|---|
| Assamese | 2,620,000 | 0 | 2,620,000 |
| Bengali | 5,520,000 | 442,000 | 5,962,000 |
| Gujarati | 12,150,000 | 564,000 | 12,714,000 |
| Hindi | 12,390,000 | 588,000 | 12,978,000 |
| Kannada | 2,240,000 | 0 | 2,240,000 |
| Kashmiri | 2,270,000 | 0 | 2,270,000 |
| Malayalam | 2,350,000 | 0 | 2,350,000 |
| Marathi | 2,210,000 | 0 | 2,210,000 |
| Oriya | 2,730,000 | 0 | 2,730,000 |
| Punjabi | 15,600,000 | 521,000 | 16,121,000 |
| Sinhala | 6,860,000 | 0 | 6,860,000 |
| Tamil | 19,980,000 | 0 | 19,980,000 |
| Telegu | 3,970,000 | 0 | 3,970,000 |
| Urdu | 1,640,000 | 512,000 | 2,152,000 |
| Total | 93,530,000 | 2,627,000 | 96,157,000 |

The EMILLE-CIIL monolingual written corpora (MWC) come as a result of the collaboration between the EMILLE project team at Lancaster University and the Central Institute of Indian Languages (CIIL), Mysore. The MWC corpora have a total size of approximately 93,530,000 words. In addition, spoken data (more than 2.6 million words) was also collected for those languages with a UK community large enough to sustain spoken corpus collection (Bengali, Gujarati, Hindi, Punjabi and Urdu). As members of the South Asian minority communities in Britain were uneasy with having their everyday conversations included in a corpus – even when the data was fully anonymized – most of the speech data is context-governed speech (transcripts of radio programs from the BBC Asian Network), though the Bengali and Hindi corpora also contain small amounts of de-

[8] The discussion in this paragraph is based on Corpus Encoding Standard (http://www.cs.vassar.edu/CES).

[9] The project was funded by the UK EPSRC (Grant references GR/N19106, GR/M70735, GR/N28542 and GR/R42429/01). The corpus can be accessed at the following site: http://www.ling.lancs.ac.uk/corplang/emille.

mographically sampled speech. Table 1 gives the sizes of each monolingual corpus.

In the EMILLE Corpus, files are classified by their provenance or genre. All of them have a filename in a standard format, which consists of a series of codes chained together with hyphen characters. These codes specify the main language of the file, the source of the text, its subcategory in terms of subject matter if such information is available, and an identifying number. The name is generally of the format: *[Language]-[text type]-[Source]-[subcategory]-[identifyingNumber].txt*. In the case of sources from which text was gathered on a periodical basis (i.e. the news websites in the written corpus, the radio programs for the spoken corpus) the identifying number is a date. For other files it is simply an arbitrary unique number. The major exception to this scheme is the Sinhala written corpus, which, unlike the other languages, is organized primarily by the category into which the text falls, and secondarily by the source it was gathered from.

The EMILLE monolingual corpora are balanced, covering a number of genres and domains. Whilst these corpora vary in size, each of them contains at least two million words, typically large enough for natural language processing tasks.

On the EMILLE project we wished to develop a POS tagger for at least one of the languages covered by the project. The language we chose to focus on was Urdu. We selected Urdu for a number of reasons. Firstly, it is widely spoken in the UK, both as a first and second language, and native speakers were available to be consulted at Lancaster where this part of the project took place. Secondly, as the *lingua franca* of a multilingual community (that of South Asian Muslims) and the official language of Pakistan, Urdu has considerable political and cultural importance. Thirdly, there are a number of factors that we anticipated would make tagging Urdu more complicated than tagging any other EMILLE language. For example, the right-to-left directionality of the Perso-Arabic script in which Urdu is written and the presence of grammatical forms borrowed from Arabic and Persian, which are structurally quite distinct from Indo-Aryan forms, mean that Urdu represents a unique challenge in our data. It seemed the best course of action was to confront these problems by choosing Urdu as the language for which to develop POS tagging. The Urdu tagset was created, using the Urdu grammar of Schmidt (1999) as a basis, in accordance with the EAGLES guidelines on morphosyntactic annotation (Leech and Wilson 1999). The data in Urdu (both monolingual and parallel) was annotated with morphosyntactic analysis using the Urdu tagger developed on the EMILLE project (see Hardie 2003).

The corpus annotation research of EMILLE has also expanded to cover another form of annotation – the annotation of demonstratives – in Hindi. The annotation scheme was designed on the basis of the blueprint provided by Botley and McEnery's (2001) scheme devised for English demonstrative anaphors (see Sinha 2003 for details). The anaphorically annotated Hindi corpus contains roughly 100,000 words of news material (20 excerpts from the *Ranchi Express* data).

The parallel corpus was compiled using 75 advice leaflets published by the UK government, taking the form of approximately 200,000 words of English originals with accompanying translations in five South Asian languages (Hindi, Bengali, Punjabi, Gujarati and Urdu). The research value of these British government data is very high in our view. Whilst the corpus is composed of only one genre, it covers a range of domains, including consumer issues, education, housing, health, law and social security, all of which are term-rich areas.

The EMILLE Corpus is a product of collaboration between the Lancaster team and Central Institute of Indian Languages (CIIL), Mysore, India. We learnt from our experience that collaboration is better than competition. The construction of large-scale language resources needs to accept this truth if it is to be effective (see Baker et al 2003).

## 4 The LCMC Corpus

The Lancaster Corpus of Mandarin Chinese (LCMC) is a one million word balanced corpus of written Mandarin Chinese. The corpus was created as part of the research project *Contrasting tense and aspect in English and Chinese*.[10] We built the LCMC Corpus in response to the general lack of publicly available balanced corpora of Chinese (see section 1). The only balanced corpus of Mandarin Chinese is the Sinica Corpus, which was produced by Academia Sinica, Taiwan. As a result

[10] The project was funded by the UK ESRC (Grant reference RES-000-220135).

of Taiwan being separated politically from Mainland China for decades, the language used in Taiwan has diverged from that used in Mainland China.[11] As such, the Sinica corpus does not represent modern Mandarin Chinese as written in Mainland China. Given the available corpus resources for Chinese corpus linguistics and our desire to use a balanced corpus of modern Mandarin Chinese from Mainland China to contrast English and Chinese, we decided to build LCMC.

Table 2: Genres covered in the LCMC Corpus

| Code | Text category | Samples | Proportion |
|---|---|---|---|
| A | Press reportage | 44 | 8.8% |
| B | Press editorials | 27 | 5.4% |
| C | Press reviews | 17 | 3.4% |
| D | Religion | 17 | 3.4% |
| E | Skills, trades and hobbies | 38 | 7.6% |
| F | Popular lore | 44 | 8.8% |
| G | Biographies and essays | 77 | 15.4% |
| H | Miscellaneous (reports and official documents) | 30 | 6% |
| J | Science (academic prose) | 80 | 16% |
| K | General fiction | 29 | 5.8% |
| L | Mystery and detective fiction | 24 | 4.8% |
| M | Science fiction | 6 | 1.2% |
| N | Adventure and martial arts fiction | 29 | 5.8% |
| P | Romantic fiction | 29 | 5.8% |
| R | Humor | 9 | 1.8% |
| Total | | 500 | 100% |

As the LCMC Corpus was created principally with contrastive research in mind, it was designed as a Chinese match for the FLOB (British English, see Hundt, Sand and Siemund 1998) and Frown (American English, see Hunt, Sand and Skandera 1999) corpora. All three corpora sampled written text produced in 1991 – 1992, covering 15 genres as shown in Table 2.

LCMC has been constructed using written Mandarin Chinese texts published in Mainland China to ensure some degree of textual homogeneity. It should be noted that the corpus is composed of written textual data only, with items such as graphics and tables in the original texts replaced by <gap> elements in the corpus texts. Long citations from translated texts or texts produced outside the sampling period were also replaced by <gap> elements so that the effect of translationese could be excluded and L1 quality guaranteed.

While a small number of samples, if they were conformant with our sampling frame, were collected from the Internet, most samples were provided by the SSReader Digital Library in China. As each page of the electronic books in the library came in PDG format, these pages were transformed into text files using an OCR module provided by the digital library. This scanning process resulted in a 1-3% error rate, depending on the quality of the picture files. Each electronic text file was proofread and corrected independently by two native speakers of Mandarin Chinese so as to keep the electronic texts as faithful to the original as possible.

As we needed to the follow the structure of the FLOB/Frown corpora, all of the samples for each genre were combined into one file. As such, in building the LCMC Corpus, we had to modify the CES header (see section 2) slightly by moving the bibliographic information for each sample into a separate ancillary document accompanying the corpus.

While the original data was encoded with GB2312, we decided to convert the corpus into Unicode, following the standards established in section 2. In addition to the standard version containing Chinese characters, we also produced a Pinyin version to enable users who can read Romanized Chinese but not Chinese characters to use our corpus.

The corpus is XML-conformant. Each file has two parts: a corpus header and the corpus text itself. The header contains general information about the corpus. The text part is annotated with five main features: text category (genre), sample file, paragraph, sentence and token. We undertook two forms of corpus annotation on the LCMC Corpus: word segmentation and part-of-speech annotation. Automatic processing achieved a precision rate of 97.16% for POS tagging, which was improved to over 98% by post-editing (see McEnery, Xiao and Mo 2003).

[11] In Taiwanese Mandarin, for example, *you* can function as a perfective marker indicating the actualization of a situation, especially in conversations. Speakers of Mainland Mandarin find this usage odd and even ungrammatical (cf. Christensen 1994).

The LCMC Corpus is a valuable resource for research into Mandarin Chinese and, in combination with FLOB and/or Frown, for the contrastive study of Chinese and English. It is our hope that the release of LCMC will stimulate corpus-based research both into modern Chinese itself, and into modern Chinese in contrast with English.[12]

## 5 Exploration tools

As EMILLE and LCMC are marked up respectively in SGML and XML, non-markup-aware concordancers will not allow users to easily exploit these corpora fully. Two Unicode-compliant markup-aware corpus tools that are available, *Xara* (Burnard and Todd 2003) and *WordSmith* version 4 (Scott 2003), are at the final stage of beta testing at the moment and will be released soon.

Using *WordSmith* 4 to explore the two corpora is quite straightforward, though the LCMC Corpus needs to be converted from utf-8 to utf-16 first using a built-in utility of *WordSmith*.

*Xara* is more powerful in that it allows users to build very complex queries, yet it is accordingly more difficult to use. The program is an XML-compliant extension of *SARA* (SGML-aware Retrieval Application) originally developed for the British National Corpus (cf. Aston and Burnard 1998). It can be used for both the local and remote access of a corpus. With *Xara*, a corpus needs to be indexed using the Indexer tool before it can be explored using the client program. This section demonstrates how to explore EMILLE and LCMC using *Xara*. We will also introduce a web-based concordancer developed for LCMC.

When we indexed the EMILLE Corpus, we followed the corpus architecture established in section 3. The 14 monolingual corpora (including speech data where appropriate) were indexed separately. The two annotated corpora for Urdu and Hindi were also kept separate, as they consist of data contained within the monolingual corpora. Similarly, the six parallel corpora were indexed individually by language. As a result, there are altogether 22 subcorpora in EMILLE. Note, however, that using *Xara* to explore these parallel corpora is an interim solution. We are aligning these corpora at the sentence level, using an alignment algorithm developed at Lancaster (see Piao 2000). Once the alignment is complete, users will be able to explore these parallel corpora with the new Unicode-compliant version of *ParaConc* developed by Michael Barlow, which is now being beta tested.

As the EMILLE Corpus is richly annotated with various kinds of information, there are many different markup elements in the corpus. The most important parameters which are directly relevant to most users include *channel*, *domain* (for monolingual written corpora), as well as *occupation* and *person* (indicating a speaker's occupation, sex and age for speech data). POS tags and anaphoric tags are also important, respectively, for annotated Urdu and Hindi data. These two types of token-level tags were indexed following a different policy from other elements to allow for POS queries. In the remainder of this section, we will use LCMC to demonstrate how to explore these corpora with *Xara*. While the languages discussed are different, the method of using *Xara* on each should remain the same.

In the LCMC Corpus, the most important XML elements are *text* (text category), *file* (sample file), *s* (sentence) and *w* (word token).[13] The *text* element can be used to compare different genres while the *file* and *s* elements indicate the location of a concordance to provide a reference back in the corpus. Now suppose we want to extract all instances of the verbal-final 了 *-le* (tagged as *u*) immediately followed (the link type defined as *Next*) by a noun (tagged as *n*) in sentence number 0010 in all of the 500 sample files in the 15 text categories. This complicated query can be made using "Query builder" of *Xara*. First, define the *scope node* (the left node in Query builder that indicates the context to search in) as "0010" using the *s* element (Fig. 1). In the *query node* (the right node in Query builder), select *AddKey* (POS) to define the first part of the query as 了 and select the POS tag *u*, and the second part as *Any* and select the POS tag *n*. Then define the *link type* as *Next* (Fig. 2). The search result is shown in Fig. 3. The upper part of the concordance window gives the query text (Select *Query – Query text* from the main menu to display

[12] See McEnery, Xiao and Mo (2003) for a discussion of the design criteria and technical details of the corpus. The corpus is can be accessed at the following website: http://www.ling.lancs.ac.uk/corplang/lcmc.

[13] Following the BNC style, punctuations and symbols in LCMC are tagged separately from word tokens using the *c* element.

the query text) while the lower window displays the concordances. The status bar of the concordance window shows the name of the corpus, the current position of the pointer/mouse (i.e. concordance number 1), the total number of concordances (i.e. 25), the number of files in which the query is matched (10), the file name where the current concordance occurs (i.e. LCMC_A), and the file/sentence number for the current concordance (i.e. File A04 and sentence number sn0010). As we have searched in sentence number 0010 (in 500 sample files), this should be the sentence number for all of the concordances.

Figure 1: Defining the scope node

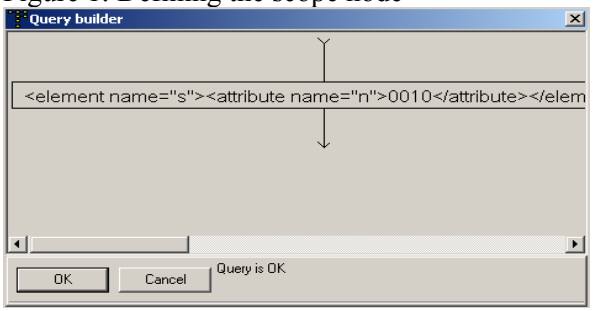



Figure 2: Defining the query node

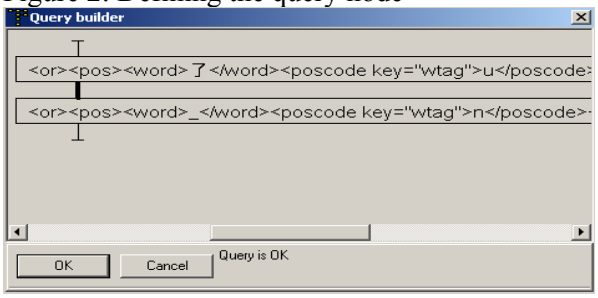
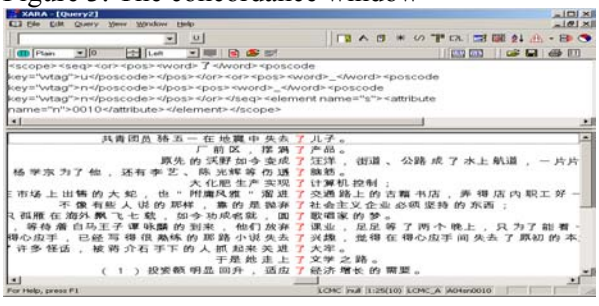



Figure 3: The concordance window



By comparison to many other corpus tools, one advantage of *Xara* is that it displays complete sentences while also centering the search query. Users are also given options to display concordances in the *page* (giving more context) or *line* mode (i.e. KWIC, as shown in Fig. 3), in XML or plain text. Additionally, users can define their own style sheet to display selected XML elements. *Xara* can also compute significant collocates automatically using a statistic selected from those available by the user.

While the EMILLE and LCMC corpora can be explored most efficiently with *Xara*, we have also developed a web-based concordancer (*WebConc*) for use with LCMC, which is more user-friendly than *Xara*.

*WebConc* allows users to search in the standard character version or the Romanized Pinyin version of the LCMC corpus using a token, POS tag or their combination. Users can also select text categories for inclusion in their search. The search result can be displayed the sentence or KWIC mode (both displaying complete sentences), in XML or plain text. The concordancer also gives a summary of the query, including the query text, the date the corpus is accessed, raw and normalized (per million words) frequencies in each text category, and the total frequency in the text categories users have selected. Together with the instructions for ordering LCMC, *WebConc* can be accessed at the corpus website given in section 4.

## 6 Conclusion

This paper has discussed standards for developing Asian language corpora and presented two corpora developed at Lancaster University, together with exploration tools for use with these corpora. The standards we propose here work well with Asian language corpora, as demonstrated by our practice in corpus development; they also conform to the current trends in the international NLP community. The two corpora we developed also constitute an improvement to the previous state of Asian language resources.

Asia is a continent of many languages and is potentially rich in language resources. The creation of Asian language resources is growing, as witnessed by the first three ALR-workshops. However, there is still a lot to be done. One thing that may ease the current situation is a true Asian parallel to ELRA that can coordinate corpus development efforts in the region. The situation could also be improved by corpus builders working on Asian languages standardizing corpora so as to facilitate data interchange. We have learned from our work with CIIL on the EMILLE project that collabora-

tion is better than competition. Our experience in collaborating with the *Xara* team also tells us that the cooperation between corpus creators and software developers can produce better corpora and better corpus tools. It is our belief that the cooperation and collaboration between centres and institutes worldwide will undoubtedly give rise to the further development of Asian language corpora.

## References


Aston, G. and Burnard, L. 1998. *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Baker, P., Hardie, A., McEnery, A., Cunningham, H., and Gaizauskas, R. 2002. "EMILLE, a 67-million word corpus of Indic languages: data collection, mark-up and harmonization". In *LREC 2002 Proceedings*, pp. 819-827.

Baker, P., Hardie, A., McEnery, A. and Jayaram B. 2003. "Corpus data for South Asian language processing". In P. Hall and D. Rao (eds) *Proceedings of the Workshop on Computational Linguistics for the Languages of South Asia*, pp. 1-8.

Botley, S. and McEnery, A. 2001. "Demonstratives in English: a corpus-based study". *Journal of English Linguistics*. 29: 7-33.

Burnard, L. and Todd, T. (2003). "Xara: an XML aware tool for corpus searching". In *Proceedings of Corpus Linguistics 2003*, pp. 142-4.

Christensen, M. 1994. *Variation in Spoken and Written Mandarin Narrative Discourse*. Ph.D. thesis, Ohio State University, Columbus.

Goto, I., Kato, N. and Ehara T. 2001. "A multilingual news database and its application to a translation memory system". In *Proceedings of ALR-2 Workshop*, pp. 1-6.

Gu, Y. 2002. "Towards an understanding of workplace discourse". In C. Candlin (ed) *Research and Practice in Professional Discourse*, pp. 137-86. Hong Kong: City University of Hong Kong Press.

Hardie, A. 2003. "Developing a model for automated part-of-speech tagging in Urdu". In *Proceedings of Corpus Linguistics 2003*, pp. 298-307.

Huang, C. 2001. "Current Status and Future of Language Resources in Taiwan". In *Proceedings of ALR-1 Workshop*.

Hundt, M., Sand, A. and Siemund, R. 1998. *Manual of information to accompany the Freiburg - LOB Corpus of British English ("FLOB")*.

Hunt, M., Sand, A. and Skandera, P. 1999. *Manual of information to accompany the Freiburg - Brown Corpus of American English ("Frown")*.

Ide, N. 2002. "Requirements, tools, and architectures for annotated corpora". In *LREC 2000 Proceedings of Data Architectures and Software Support for Large Corpora Workshop*, pp. 1-5.

Leech, G. and Wilson, A. 1999. "Standards for tagsets". In H. van Halteren (ed) *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers.

McEnery, A., Baker, P., Gaizauskas, R. and Cunningham, H. 2001. "EMILLE: building a corpus of South Asian languages". In D. Lewis and R. Mitkov (eds) *Proceedings of Machine Translation and Multilingual Applications in the New Millennium*.

McEnery, A., Xiao, Z. and Mo, L. 2003. "Aspect marking in English and Chinese: using the Lancaster Corpus of Mandarin Chinese for contrastive language study". *Literary and Linguistic Computing* 18(4): 361-78.

McEnery, A. and Xiao, Z. 2004. "Character encoding in corpus construction". In M. Wynne (ed) *Guide to Good Practice*. Oxford: Oxford University Press.

Piao, S. 2000. *Sentence and Word Alignment between Chinese and English*. Ph.D. thesis, Lancaster University, Lancaster.

Rim, H. 2001. "Language Resources in Korea". In *Proceedings of ALR-1 Workshop*.

Schmidt, R. 1999. *Urdu: an essential grammar*. London: Routledge.

Scott, M. 2003. *WordSmith Tools Manual*.

Shirai, K. 2001. "Language resources in Japan". In *Proceedings of ALR-1 Workshop*.

Sinha, S. 2003. *Demonstrative Anaphors in Hindi Newspaper Reportage: A Corpus-based Study*. MA dissertation, Lancaster University, Lancaster.

Sornlertlamvanich, V. 2001. "Thai linguistic resources". In *Proceedings of ALR-1 Workshop*.

Thongprasirt, R., Charoenporn, T., Sinthupinyo, W. and Sortlertlamvanich, V. 2001. "Development of very large corpora in Thailand". In *Proceedings of ALR-2 Workshop*.

Wittenburg, P., Peters, W. and Broeder, D. 2002. "Metadata proposals for corpora and lexica". In *LREC 2002 Proceedings*, pp. 1321-6.