Toward a Digital Library with Search and Visualization Tools

John Lee, Ying Cheuk Hui, and Chak Yan Yeung

Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
{jsylee, yingchui, chayeung}@cityu.edu.hk

Abstract

We present a digital library prototype with search and visualization capabilities, designed to support both language learning and textual analysis. As in other existing libraries, users can browse texts with a variety of reading aids; in our library, they can also search for complex patterns in syntactically annotated and multilingual corpora, and visualize search results over large corpora. With a web interface that assumes no linguistic or computing background on the part of the user, our library has been deployed for pedagogical and research purposes in diverse languages.

Keywords: digital libraries, corpus search, corpus visualization, parallel corpora, treebanks

1. Introduction

In the past decades, language learners have increasingly made use of digital texts. These texts have been compiled and displayed in many digital libraries, now covering a wide range of domains and languages, from Latin to ancient Greek\(^1\) to Classical Chinese\(^2\). These libraries provide an environment for users to not only browse the texts, and also to access reading aids such as morphological analyses, dictionary entries, and contextual information. More recently, many digital texts have also been richly annotated with multilingual information, including sentence alignments (e.g., Erjavec et al., 1998) and word alignments (e.g., Lee et al., 2012); as well as linguistic information, such as parts-of-speech and syntactic structure (e.g., Haug & Jøndal, 2008).

These annotations are certainly valuable as additional reading aids, to help the reader better understand individual sentences better. They can, however, also open up new methods of learning and analysis, by revealing linguistic patterns — for example, the different translations or selectional preferences of a word (e.g., Bamman & Crane, 2008), or the distribution of a syntactic structure (e.g., McEnery et al., 2003) — not over individual sentences, but over the entire text corpus.

To exploit these new methods, a digital library must enable users not only to browse, but also to query the text corpus and its annotations. For simple queries, such as keyword search, a relatively straightforward interface suffices. Complex linguistic annotations allow queries about more complex patterns, such as those suggested above; however, these queries are inevitably more difficult for humans to communicate to the machine.

To enable non-technical users to formulate and pose these queries, a suitable user interface is essential. Interface design in linguistic computing has not received much attention, for practical reasons. Until recently, corpora with multilingual or linguistic annotations tended not to be read by humans but by machines, serving as training data for natural language processing tasks such as machine translation and parsing. As these corpora began to be exploited by humanities scholars, many built search interfaces through the use of query languages (e.g., Carletta et al., 2003); these languages are powerful and precise, but tend to demand a steep learning curve for users without substantial technical competence.

Despite these advances, search and visualization tools remain underused by humanities scholars, let alone the general public. In this paper, we describe a digital library that seeks to lower the barrier for all to benefit from these tools. The user can browse complex linguistic data, including multilingual information and syntactic annotations; perform sophisticated search on these annotations; and interpret the search results with visualization tools. Requiring minimal computing or linguistics skills, the library has been deployed in various classroom and research settings, in multiple languages, for users with no technical background.

The next section describes the interface of the digital library; an outline of each of the three environments is followed by a brief summary of its use, and a review of relevant previous work. Section 3 provides technical details on the implementation of the library. Section 4 concludes and outlines future work.

2. Interface Description

Our digital library offers three environments, among which the user can easily toggle. Each environment facilitates a different activity. The ‘Browse’ environment (section 2.1) is designed for reading sentences sequentially; the ‘Search’ environment (section 2.2) is intended for retrieving examples of words or linguistic patterns from the entire corpus; the ‘Visualize’ environment (section 2.3) graphically displays search results to help reader interpret their significance.

2.1. Browse

In the ‘Browse’ environment, upon choosing one or more corpora from a list, the user sees sentences from the corpus displayed in order. He or she can then view translations and linguistic information for each sentence.
Fig. 1: The ‘Browse’ environment supports bilingual reading. The original Greek sentences from the beginning of Bibliotheca are shown on the left, and their Portuguese translations on the right. When mousing over a Greek word, such as Βριαρεύς (‘Briareus’, name of a Greek god), its Portuguese gloss (‘Briarão’) is shown in a pop-up text box, and its equivalent word in the Portuguese translation (‘Briareu’) is highlighted.

**Bilingual information.** For parallel, sentence-aligned corpora, the user can simultaneously read sentences in two languages, displayed side by side. Figure 1 shows an example with a Greek-Portuguese parallel corpus of Bibliotheca, an ancient Greek collection of myths and legends, which has been translated into Portuguese and word-aligned. In addition, each word is annotated with a Portuguese gloss. When the user mouses over a Greek word, the word is highlighted and its gloss is displayed in a pop-up box; the equivalent Portuguese word in the translation is also highlighted. These features are intended to help students access more advanced material early in their study, without the need to frequently turn to dictionaries.

**Applications.** The Greek-Portuguese parallel corpus of Bibliotheca is being prepared for teaching Greek to Portuguese-speaking students. The ‘Browse’ environment was deployed at an Ancient Greek course at City University of Hong Kong, for students to access the Greek-Chinese Interlinear of the New Testament Gospels.

2.2. Search

2.2.1. Keyword search

The simplest form of search is keyword search; i.e., retrieve all sentences that contain a particular word or set of words, typically displayed in the keyword-in-context format. In the ‘Search’ environment, the user is initially presented with a single text field, with a virtual keyboard (http://alllanguages.info) to aid the input of non-ASCII text. As shown in Figure 3, the user can insert as many text fields as desired to enter multiple keywords. The user can also generalize the search to any surface form of the keyword by choosing “any word form”, or to any word of a particular POS by entering the POS tag instead.

**Applications.** The search can be combined with bilingual browsing (see section 2.1). In a Cantonese class for Mandarin speakers, students were given a list of polysemous Mandarin words — each word has two or more possible Cantonese translations, depending on context. Using the word-aligned Cantonese-Mandarin parallel corpus (Lee, 2011), students searched for these Mandarin words, and then examined the possible Cantonese translations. Results of a post-test suggested that this exercise significantly raised the students’ mastery of the use of these Cantonese words.

2.2.2. Relation between words

Beyond the look-up of independent words or n-grams, one is often interested in retrieving examples of multiple words that are related in some way. These words may be non-adjacent, or even reside in different corpora. For example, one might wish to investigate which words can serve as the noun subject of the verb ‘to love’; or to explore the nuances of the meaning of ‘to love’, by retrieving its equivalents in other languages. Users can formulate these queries by first entering the words or POS concerned (2.2.1), then further constraining them by adding relations between them. A word may have one of two types of relations with another word: it can modify another word, typically in the context of a dependency treebank; and it may be aligned to another word, typically in the context of a parallel corpus.
 ALIGNMENT RELATIONS. A second kind of relation is word alignment in parallel corpora. Most commonly, these alignments represent translations, but they can also indicate other relations such as revision (e.g., Daxenberger & Gurevych, 2012). These alignments can restrict search results to keywords in a particular sense. Another use is to retrieve all possible translations of a word. In Figure 3, the user searches for all Greek verbs meaning ‘to love’ by specifying that the Chinese word ai ‘to love’ must be aligned to the verb (‘V-’).

A convenient feature of this interface is the capability to mix and match an arbitrary number of these two types of relations. For a Chinese speaker who wishes to retrieve all sentences in New Testament describing the ways a father ‘loves’, the query would run as in Figure 3: a verb is required to be aligned to the word ai ‘love’; further, the word “father” is to be the subject of the main verb. A snippet of the search results, showing two different such verbs, are shown in Figure 4.

APPLICATIONS. This search interface has been deployed in an Ancient Greek course at City University of Hong Kong for Chinese-speaking students. Instead of learning about various Greek adjectival constructions and verb endings from a textbook, students searched for examples observed the patterns (Lee et al., 2012). In another course, students used this interface to analyze the use of colors (2.3), pivot constructions and other grammatical phenomena in Classical Chinese (Lee et al., 2013).

PREVIOUS WORK. Query languages are the most common interface for searching treebanks. Tgrep2 (Rohde, 2001), for example, is the standard language for the Penn Treebank. Other systems have their own variants (Carletta et al. 2003; Zeldes et al. 2009). Query languages have the advantage of succinctly expressing a wide range of precise queries. However, users, especially those without programming skills, often struggle with learning the syntax of these technical languages.

A common strategy to deal with this complexity is to use a graphical query builder. Instead of typing out a terse query, the user draws a partial graph or subtree, with search nodes and operators on edges linking these nodes. Some recent examples include TIGERSearch (Lezius, 2002) and Netgraph (Mirovský, 2006). Some of these interfaces, like ours, also support simultaneous search on syntactic and alignment relations (e.g., Nyaarda & Johannessen, 2004; Volk et al., 2007).

For users without a background in linguistics, it can be intimidating to draw parse trees and to populate its nodes and edges. Our search interface has similar power of expression and, in our experience, students and other non-experts in various projects were able to master it with little instruction or training (Lee et al., 2013).

2.3. VISUALIZE

When the volume of search results is large, the user may have difficulty seeing the significance or broader patterns in the keyword-in-context format offered in the ‘Search’ environment. Depending on the research question, the user might wish to see how frequently, and where, certain linguistic constructions are located. The ‘Visualize’ environment offers two modes to do so.

WORD DISTRIBUTION. In this mode, the entire corpus is abstracted as a sequence of black dots. One dot represents either one word, or one sentence, depending on the scale desired by the user; s/he can then create a picture highlighting particular keywords, which can be words, lemmas or parts-of-speech. Wherever the keyword occurs, the corresponding dot turns from black to a customizable color, thus showing at a glance the positions and the frequency of multiple keywords. Figure 5, for example, shows graphically how two synonymous terms dominate different parts of a corpus.

NETWORKED RELATIONS. Word frequencies can be sorted and listed in a table, or graphically represented by different font sizes in a word cloud. With the capability to search for syntactic relations between words (2.2.2), one may similarly display frequencies of these relations.

The use of vivid colors is an important literary device in Classical Chinese poems. One can visualize, for example, the colors most frequently ascribed to ‘mountains’. More generally, one might want to further investigate the sets of objects that overlap in terms of their colors; in other words, to retrieve the sets of colors that are commonly shared by a set of objects. This query can be a starting point of research into, for example, the kinds of colors most typically used in describing objects of nature.
The user can specify a set of dependency relations (e.g., adjectival modifier), a set of keywords as children (e.g., color words) and another set of keywords as head words (e.g., objects of nature). A one-dimensional list would not suffice to visualize the results; instead, we interpret them as a set of “networked” relations, and represent them as a graph, using the Graphviz library. The nodes represent the keywords; an edge means that the keywords are connected with one of the specified relations. Figure 7 shows an example graph on the use of colors.

Previous Work. There are many variants for visualizing word distributions in terms of their locations, including Voyant Tools (voyant-tools.org), TextArc (textarc.org), Arc Diagrams (Wattenberg, 2002), and Blue Dots (Lancaster, 2010). None of them, however, has incorporated syntactic relations, and cannot display networked relations.

Applications. In a course on language information technology, students constructed graphs to visualize various phenomenon in Classical Chinese, including the use of colors in poems (Lee et al., 2013), and the most common verbs associated with principle characters in selected novels.

3. Implementation

As shown in Figure 6, our digital library has three components, a web interface, a relational database, and a web service. The web interface uses a client-side JavaScript framework, jQuery, and a UI library called jQueryUI. From keywords entered by the user, the system generates an SQL query for searching. We have chosen to use a relational database, MySQL, rather than XML files, to avoid cross-referencing entities in different files and to speed up search time. Figure 6 shows a entity-relation diagram of our schema. The ‘Word’ table stores all word tokens, as well as their associated linguistic features, such as lemma, part-of-speech and dependency label, all of which are stored in foreign key fields with indices. Pairs of aligned words, mostly translations of one another, are stored in a cross-reference table called “Alignment”. Higher up in the hierarchy, various tables store metadata at the sentence- (e.g., verse number), chapter-, book- and corpus levels. Finally, for client-service communication, we adopt XML-RPC.

Fig. 4: Partial results of a search query in Figure 3, combining word alignments and dependency relations. Highlighted on the left are the various Greek verbs meaning ‘to love’, of which the word ‘father’ (πατήρ) is attested to be the agent. On the right are the Chinese translations, with the word ‘to love’ highlighted.

Fig. 5: Visualization of word distribution: the use of the terms “God” and “LORD”, two names for the same entity in the book of Genesis. Each dot represents one sentence; the dot is coloured red if the sentence contains the word “God”, and yellow if it contains “LORD”. The picture shows that the first term dominates the beginning part of the book, but eclipsed by the second term in the later parts.

Fig. 6: On the left, the diagram illustrates the overall structure of our system. On the right, the entity-relation diagram shows the schema of the relational database that stores linguistic data.
4. Conclusion and future work

We have described a digital library designed for language learning and text analysis. It supports bilingual browsing, searching in richly annotated corpora, and visualization of search results.

This library is intended to lower the threshold for non-technical users to work with massive text corpora with structured annotations, so as to discover patterns that are difficult to detect with traditional reading methods. It has been validated for both pedagogical and research purposes, in university courses and research projects in various languages, including both modern and Classical Chinese, Portuguese, and ancient Greek.

We plan to further simplify our search interface with example-based approaches (cf., Augustinus et al., 2012), and to create customizable visualization modules.

Acknowledgments

This project was supported in part by a grant from the Early Career Scheme of the General Research Fund (#9041849) from the Research Grants Council of Hong Kong, an Interdisciplinary Professional Development Award (#9088003) and a Teaching Development Grant (#7002549) from City University of Hong Kong.

References


