A Corpus of Commented Editions of Literary Texts

John Lee[1], *Caio Camargo*[2], *Yin Hei Kong*[1]

(1) City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
(2) Sao Paulo State University (UNESP), Rod. Araraquara-Jaú, Km.1, Machados, Araraquara-SP, Brazil

jsylee@cityu.edu.hk,caiocamargo1986@gmail.com,yhkong@cityu.edu.hk

ABSTRACT

Study of a literary text requires an understanding of its original context, which can be gleaned from other related texts. Humanities scholars often gain pointers to these relevant texts from the so-called "commented editions", which provide citations of other texts that might have served as source material or can provide background information. This paper presents a corpus that contains more than 1,100 such citations drawn from text editions of literary works in Ancient Greek and Classical Chinese, and a collection of 9-million-word texts from both literary traditions. The corpus is intended to facilitate the development and evaluation of semantic search methods to automatically generate commented editions for literary texts.

KEYWORDS: corpus, citations, commented editions, Ancient Greek, Classical Chinese.

1 Introduction

Accurate and thorough interpretation of a literary text requires not only a close examination of its own content, but also knowledge of its original context. The context is typically gleaned from related texts, such as those that might have served as source material, or might provide useful contemporary information. To avail themselves of pointers to these relevant texts, humanities scholars often consult the so-called "text editions" or "commented editions", which provide not only the raw text but also comments, in the form of footnotes or the more formal "critical apparatus". A comment usually cites one or more passages from other texts to shed light on a wide range of issues: they may explain the unusual spelling of a word or give variant readings of a sentence; provide background information on a geographic location; or offer a passage in another text as possible source material. Table 1 shows two example comments, drawn from Frazer's (1921) text edition of *Bibliotheca*, a handbook of Ancient Greek mythology.

Producing these text editions is a monumental effort requiring much time and encyclopedic knowledge. Not surprisingly, only the more popular literary works have received the publishers' attention; the vast majority of classical texts still await their own commented editions.

Recent advances in natural language processing may now be applied towards automatic generation of such commented editions. This task can be viewed as a generalization of text reuse detection (Bamman & Crane, 2008; Buechler et al., 2010), to include other related passages that are not paraphrases but provide relevant background information. This paper takes a first step towards this goal by presenting a corpus for developing and evaluating semantic search algorithms for this purpose. If successful, these editions would be a boon to the study of less popular texts, but also has the potential to discover related passages yet unnoticed by the canonical editions.

2 Corpus

Our corpus consists of texts, encoded in TEI-compliant XML, and citations, in stand-off annotation. We describe the target texts (section 2.1), the citations (section 2.2), and finally the candidate texts (section 2.3).

2.1 Target Texts

'Target texts' refer to those texts in our corpus to which the comments are addressed. Our corpus contains target texts in both prose and poetry from two major literary traditions, Ancient Greek and Classical Chinese. Table 2 gives the list.

Ancient Greek. *Bibliotheca*, traditionally attributed to Apollodorus, is a 27K-word handbook of Greek mythology composed around 200 BCE. Frazer's (1921) commented edition is considered the foremost for this widely studied book. This work is divided into three books and 209 sections, with an average length of about 130 words each. These sections are not the most natural unit of analysis for our purpose since each section receives an average of almost four comments, with citations focusing on different aspects. We have therefore further divided each section into sentences.

Classical Chinese. The target texts are drawn from the works of Wang Wei and Du Fu in the *Complete Tang Poems*. Often considered two of the greatest Chinese poets, both Wang and Du lived in the 8th century CE. Our corpus is based on a subset of their poems, containing about 8,600 characters, with comments and citations given in the definitive editions by Zhao (1736) and Chou (1764).

<table>
<tr><td colspan="3">Sky was the first who ruled over the whole world … After these, Earth bore him the Cyclopes, to wit, Arges, Steropes, Brontes, of whom each had one eye on his forehead. <b>[1]</b> But them Sky bound and cast into Tartarus, ... <b>[2]</b></td></tr>
<tr><th>Comment</th><th>Annotation</th><th>Cited text</th></tr>
<tr><td>[1] Compare <b>Hes. Th. 139ff</b>.</td><td>Type: <code>source</code>
Target text span:
<code>Bibliotheca 1.1.2</code>
Cited text span:
<code>Hes. Th. 139—145</code></td><td>… And again, she bore the Cyclopes, overbearing in spirit, Brontes, and Steropes and stubborn-hearted Arges, who gave Zeus the thunder and made the thunderbolt: in all else they were like the gods, but one eye only was set in the midst of their foreheads.</td></tr>
<tr><td>[2] For the description of Tartarus, <b>Hes. Th. 717ff</b>.</td><td>Type: <code>background</code>
Target text span:
<code>Tartarus</code>
Cited text span:
<code>Hes. Th. 717—721</code></td><td>… as far beneath the earth as heaven is above earth; for so far is it from earth to Tartarus.</td></tr>
</table>

TABLE 1: The top row shows an example target text, in this case the beginning sentences (in English translation) of *Bibliotheca*. Shown below are two comments on these sentences. The left column lists the original comments in Frazer (1921), with the citations bolded (e.g., "Hes. Th. 139ff"). The middle column shows annotations in our corpus, including the citation type, the target text span and the cited text span. The right column displays the cited passages.

Target texts	*Bibliotheca*	Selections from *Complete Tang Poems*
Text editions	(Frazer, 1921)	(Zhao 1736); (Chou, 1764)
# words	27507	8629
# sentences	1491	1649
# citations	824	321

TABLE 2: Statistics on the target texts in our corpus. The candidate texts are described in section 2.3.

2.2 Citations

We harvested citations from the text editions listed in Table 2. For each citation, we annotated its type and associated text spans.

2.2.1 Citation type

Each citation is assigned as one of two types — `source` or `background` — according to the nature of the information it provides.

The source citations indicate possible sources of the sentence in the target text; comment [1] in Table 1, for example, lists "Hes. Th. 139ff" — i.e., line 139 and following in Hesiod's *Theogony* — as the source for its target sentence from *Bibliotheca*. While modern authors are expected to provide clear citations and references when they base their writing on previous sources; ancient authors, however, tend not to do so; in fact, readers were expected to know the sources.

The background citations give more details on a term in the target sentence, e.g., a geographical location or an object; the cited passage may not have any connection with the target sentence aside from the term. In Table 1, comment [2] points to lines 717 and following in Hesiod's *Theogony* as a description of the place Tartarus.

Among citations in the text edition of *Bibliotheca*, 67% are of the source type; the corresponding figure for the Classical Chinese poems is 88%. The distinction between these two types of citations is not only interesting for their own sake, but also useful in the sense that they will likely require different retrieval strategies. On the one hand, the source citations tend to point to passages with lots of overlapping words with the target sentence. On the other hand, in the case of background citations, there might only be one, or even none; a citation might point to, for example, a passage containing a word with a variant spelling.

2.2.2 Citation text spans

Each citation is associated with two text spans, one in the target text and one in the cited text. Both text spans are ambiguously marked in some text editions.

For source citations, the target text span can generally be taken to be the entire sentence preceding the footnote marker. For background citations, the target text span is usually only the one or two words that constitute the term for which background information is provided (e.g., "Tartarus" in comment [2] in Table 1); the rest of the sentence may not be relevant.

The cited text span is not completely specified when the citation is appended with "ff" i.e. "and the following" (e.g., "139ff" means line 139 and an unspecified number of the subsequent lines). In each of these cases, we specify the end point of the text span (e.g., line 145 for the cited text span in comment [1] in Table 1).

2.3 Candidate texts

The corpus must naturally include all texts that are mentioned in the citations; they would serve in measuring the recall of algorithms for automatic retrieval of citations. To measure the precision, however, the corpus needs to include also a much larger set of texts that can plausibly be cited; these are referred to as 'candidate texts'.

All candidate texts are annotated with their date of composition; this information is crucial since earlier works tend to be favored in citations.

For the Ancient Greek portion of the corpus, the candidate texts consist of all 317 books, with a total of almost 4 million words, from the Greek materials at the Perseus Project (perseus.tufts.edu), which is the largest source of online open-domain ancient Greek texts. For the Classical Chinese portion, the candidate texts contain 5.1 million characters in 63 books, which include all Classical Chinese books from the Gutenberg Project (gutenberg.org) composed during the Tang Dynasty or before.

3 Preliminary analysis

Although the corpus is designed for long-term impact as a testbed for automatic generation algorithms for commented editions, it can already provide citation patterns that quantify some current literary hypotheses.

Ancient Greek. Counting only the source citations, the author of *Bibliotheca* consulted as his sources no fewer than 55 books, an unusually large number. It is generally held that the author heavily reused material from Hesiod's *Theogony* in Book 1, but increasingly drew from other works such as Homer's *Iliad* and Pausanias' *Description of Greece* in Books 2 and 3. These claims are confirmed in our corpus.

The author of Bibliotheca was not only heavily dependent on *Theogony* in Book 1, but also largely preserved the order of the material. This makes sense as the book traces the creation of the world and its early history. One exception is the account of the flood, which is placed out of order to describe the origins of a different clan. We visualize these patterns by plotting locations in *Bibliotheca* with references to *Theogony*.

Classical Chinese. In terms of philosophy, Wang Wei is often said to be primary influenced by Daoism, and Du Fu by Confucianism. We investigated this claim by analyzing the genres of the books in their respective citations. After dividing the candidate texts into major genres, we found that Wang cited twice as much from Daoism material as from Confucianism material (23% vs. 12%), while the relative figures for Du are reversed (14% vs. 18%).

4 Conclusion and Outlook

We have introduced a corpus designed to facilitate a new semantic search task — automatic generation of commented editions of literary text. The corpus, based on commented editions of Ancient Greek and Classical Chinese texts and augmented with annotations on citation types, text spans and composition dates, aims to be a benchmark for comparing semantic search methods for the task.

References

Bamman, D. and Crane, G. (2008). The Logic and Discovery of Textual Allusion. *Proc. LaTeCH: Language and Technology for Cultural Heritage*.

Buechler, M., Gessner, A., Eckart, T., and Heyer, G. (2010). Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* **1**(2).

Chou, Z. 仇兆鰲 (1764). 杜詩詳注 [A detailed commentary of Du Fu's poems], in Ji Yun et al. (Eds.), 欽定四庫全書 [Complete library of the four treasuries].

Frazer, J. G. (1921). *Apollodorus, The Library, with an English Translation by Sir James George Frazer, F.B.A., F.R.S. in 2 Volumes*. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd.

Zhao, D. 趙殿成. (1736). 王右丞集箋注 [A commentary of Wang Wei's works], in Ji Yun et al. (Eds.), 欽定四庫全書 [Complete library of the four treasuries].