Bingenheimer *Lingua Sinica* (2015) 1:8
DOI 10.1186/s40655-015-0007-3

Lingua Sinica
a SpringerOpen Journal

RESEARCH Open Access

The Digital Archive of Buddhist Temple Gazetteers and Named Entity Recognition (NER) in classical Chinese

Marcus Bingenheimer

Correspondence:
m.bingenheimer@gmail.com
Temple University, Philadelphia, USA

Abstract

The identification of names and dates in larger corpora of historical texts is important for both traditional and digitally mediated research; it is part of reading as well as of exploring digital corpora. This paper is an introduction to a number of issues concerning named entity recognition (NER) for classical Chinese. In particular it introduces the "Digital Archive of Buddhist Temple Gazetteers" (http://buddhistinformatics.ddbc.edu.tw/fosizhi/), as a benchmark corpus for NER on classical Chinese and illustrates how marked-up corpora can provide answers to question that could not otherwise be addressed. The "Digital Archive of Buddhist Temple Gazetteers" is an open source and access archive of local histories of Chinese Buddhist sites. Names and dates were encoded with XML/TEI and associated with authority databases. The archive, which contains classical texts in a variety of genres, can serve as testing data for experiments in NER and POS tagging. The data is made available as part of the article.
We also show that for classical Chinese even a custom-made person name dictionary, created during the markup of the corpus, cannot in turn be used to parse the same corpus successfully without further intervention.

Keywords: Named entity recognition; Buddhist Temple Gazetteers; Names and dates in classical Chinese; Marked-up corpora; Computational analysis of historical sources

1 Background[1]

The modeling of printed into digital text has opened up novel ways of accessing and analyzing information. This has long been obvious to computational linguists who built the first digital corpora in the 60ies, that resulted in breakthrough publications such as Kucera and Francis (1967)[2]. But although the field of corpus linguistics grew quickly in the 70s, it was only after the twin revolutions of personal computing in the 80s and the Internet in the 90s that digital text started to have a significant impact on the Humanities. Now into the second decade of the 21st century, our interaction with text is almost always digitally mediated in some way or another.

In East Asia the widespread adoption of printing in the 10th century has resulted in the survival of a vast amount of texts. We consider these texts today from the twilight decades of the medium. Books as printed objects are still very much with us, but their days as the main medium for text seem counted. Based on existing trends, it is to be expected that in coming decades printing will go the way of stone inscriptions and handwriting – still in use and admired by aficionados, but much diminished in importance (Hilbert and López 2011, Rainie and Duggan 2012).

With the medium changes the way we interact with text, the ways we read and write, and extract information. The characteristic of digital text is the gap between its encoding and its presentation. Whatever is said in hearing range is accessible to our ears, whatever is written on stone or paper is accessible to our eyes or fingers. The terabytes of textual data, however, are inaccessible until a machine renders them for us in one form or another. This loss of autonomy caused by dependence on machine-mediation is generously offset by a gain in autonomy with regard to how, when and where we consume text. While analog text is generally accessed in a fashion and format chosen by the author or publisher, "users" (formerly known as readers, formerly known as listeners) have much greater influence on how digital text appears to them. This has consequences for how knowledge and information is used in society, including the academic study of ancient Chinese texts.

There is a large, largely untapped, corpus of private letters and notebooks, official communication and documents, annals and chronicles – primary sources for the study of Chinese history; using them we can discover hitherto unknown networks between people, regions and topics, if we can find ways to extract the information we need. Names are one of the most basic features through which historians structure knowledge. Names are what we remember and want to locate again in a text. That is why names are privileged in indices and encyclopedias. People, organizations, places and texts are fundamental to historical inquiry and one would be hard pressed to find a book on history that does not contain names. A list of names appearing in a text allows a guess about its provenance, domain and date - a task now approached algorithmically as "topic modeling".

The automated discovery of named entities – Named Entity Recognition (NER) – in digital text is commonly discussed in the context of natural language processing, artificial intelligence, or organizational data management. In the following I will address the prospects for NER in classical Chinese sources. In this paper I use the term "classical Chinese" as a cover term for all the many forms of literary Chinese that were used before 1919. This includes the Chinese of Buddhist translations, various forms of poetry, medical and legal texts, vernacular literature of different periods etc., and not only the more narrowly perceived "classical classical" Chinese of certain texts that the Confucian tradition has identified as a standard (i.e. the idiom of *Mencius*, *The Analects*, Sima Qian's *Records of the Historian*, etc.).

2 Named Entity Recognition - general strategies and application to modern Chinese

The framework for NER, based on which most approaches are developed, was described in a whitepaper by Chinchor et al. (1999). The authors define three subtasks for NER:

1. "Names Entities" (defining names of persons, locations and organizations), "Persons" here includes fictional characters, animals, and fictional animals (e.g. Bugs Bunny) (Chinchor et al. 1999: 11–13). "Locations" here includes stars and mythical locations, but not regions denoted only by their compass directions (e.g. the south) (Chinchor et al. 1999: 15–17). "Organizations" includes "the white house", and TV stations, but not generic terms such as "state police" (Chinchor et al. 1999: 20–22).
2. "Temporal Expressions" (dates and durations)
3. "Number Expressions" (amounts and their measures).

For classical Chinese sources organizational entities in the sense of the NER framework are of little relevance and will not be discussed here.

The most successful NER strategies work with a combination of rule-based segmentation, machine learning and a dictionary/gazetteer[3] approach (Nadeau and Sekine 2007). NER methods often rely, implicitly or explicitly, on some form of word segmentation. Word segmentation is one major area where text processing in Asian and other languages differs and segmentation methods need to be modified before they can be applied (Sproat et al. 1996). The practice of word separation in European writing was pioneered in the scriptoria of Irish monasteries in the seventh and eight centuries. It became widespread on the continent in the mid-tenth (Saenger 2000: 100). In East Asia, word segmentation with spaces was introduced in Korean only relatively recently in the 1970s. Contemporary Vietnamese writing uses spaces between each syllable, whereas word level segmentation happens in the mind of the reader. Chinese and Japanese is still written without spaces between words. Readers segment words, phrases and sentences as part of the reading process. This works fairly well for brains, but continues to be a challenge for algorithmic parsers, which demand rigorous procedures. The absence of clear criteria for distinguishing words from phrases in Chinese has been called "undoubtedly one of the most vexing problems in modern Chinese grammar" (Norman 1988: 156).

A national standard for the segmentation of modern Chinese was first published in 1993 – the 信息处理用现代汉语分词规范 *Xinxi chuli yong xiandai hanyu fenci guifan* "Contemporary Chinese language word segmentation specification for information processing" (GB13715) (Liu et al. 1994). This, however, has not solved the problem, in part, because readers do not perceive word boundaries uniformly, and the parsing during the reading process is, *horribile dictu*, not necessarily aligned with the national standard (Liu et al. 2013)[4]. Moreover, there are theoretical obstacles regarding how to segment certain word-types. Rule-based algorithms need to be given definitions of which prefixes and suffixes to include, but any linguistic definition of what constitutes a "word," and what therefore is a prefix to it, is to a degree arbitrary. In European languages this is obscured by the established convention of spaces. For Asian writing systems there exist various strategies for segmentation, which have different strengths and weaknesses (Xia 2000: 5–6). Nevertheless, as two eminent experts in the field write in a recent overview article, it is still true that "Chinese word segmentation remains a challenging topic in Chinese computational linguistics" (Huang and Xue 2012: 494).

Attempts at NER for modern Chinese during “bakeoff” competitions organized by the special interest group for Chinese Language Processing at the Association for Computational Linguistics, SIGHAN, achieved seemingly strong results. Different teams competed using different corpora and approaches. Eleven out of 43 teams reached an F-score of more than 85 on two corpora, with the best result at 91 (Levow 2006). For practical reasons the F-score translates into the percentage of correctly identified names. If such results could be applied to classical Chinese sources, we would thus be in a strong position: we could correctly identify eight or nine out of ten named entities. In spite of the optimistic report (ibid.), however, the numbers do not, or at least not yet, translate in readily applicable methods for the Digital Humanities. The actual results vary significantly between competitors and corpora; generally between F-scores of 60 and 90, with scores around 80 most prevalent. This might still be satisfying, if not all results had been obtained after training the algorithms with data from the very same corpus. Furthermore, the size of the training data was generally 5–10 times the size of the test data (ibid.). It is known that current NER systems are highly sensitive to differences between testing and training data (Huang and Xue 2012: 503), but for many forms of written Chinese, tagged test data is simply not available.

If even NER on modern Chinese relies on training the algorithm on sample sizes substantially larger than the test data, we are still a long way of distant reading our classical sources, because classical Chinese arguably poses even greater hurdles than modern Chinese[5]. Below I will outline some of the reasons why I believe this to be the case. Obviously there is the need for further empirical testing of this thesis.

3 Prolegomena regarding NER for Classical Chinese

According to Nadeau and Sekine (2007: 20) NER research tends to focus on “limited domains and textual genres such as news articles and web pages” in modern languages. Many, though by no means all, approaches to NER involve some form of word or phrase level segmentation, something that, as we have seen, is still problematic even for modern Chinese, in spite of the fact that sophisticated segmentation strategies for the analysis of modern Chinese were proposed relatively early (e.g. Huang et al. 1997). Classical Chinese, which for the purpose of this article is an umbrella term for all forms of written Chinese before 1911, is hardly ever mentioned in this context, partly because the lack of corporate and government interest, partly because the solutions developed for modern writing are not easily transferable to older forms of literary Sinitic.

As early as 1997 researchers at Academia Sinica in Taipei have started to consider how to construct a corpus of classical Chinese (Wei et al. 魏培泉等 1997), which resulted in the Academia Sinica Tagged Corpus of Old Chinese (http://old_chinese.ling.sinica.edu.tw). More recently, the segmentation of classical Chinese text has been considered by Luo et al. 羅鳳珠等 (2013). The authors suggest a strategy of term segmentation and NER for Tang and Song poetry (although mainly focusing on the Tang poets Li Bai, Du Fu and Han Yu). They point out several characteristic features of morphology and syntax that occur in classical Chinese poetry and that affect segmentation and NER. Some of the features are specific to verse and result from the need to accommodate bi-noms in rhyme and meter. Accordingly, their set of segmentation rules takes various aspects of Chinese prosody into account. As the authors note, many of these features are relevant for NER. Person names in poems are often abbreviated, combined, or used with appellatives. Alternative names for persons (courtesy, style, and taboo names, epithets etc.) are frequent, while locations too have a range of different historic or poetic names. It is not clear, whether the authors regard these issues as genre-specific with regard to NER. Clearly, most of the NER characteristics that they diagnosed for poetry (ibid., p. 5 *f*) are applicable to prose as well.

Similarly genre-specific, Xiong et al. (2013) have offered rules for segmentation and named entity annotation for Early Mandarin prose. Their corpus consists of four widely studied Ming-Qing Dynasty novels. The rule set they propose explicitly builds on existing segmentation standards for modern Chinese and adjusts these for the Early Mandarin of Ming-Qing prose, which is characterized as follows "Compared to modern Chinese, the most distinctive features of the Ming and Qing novels include frequent use of single-character words, the chapter-by-chapter style, lexical meaning, and the use of named entities, which are the key elements to be considered during the design of the specification" (Xiong et al. 2013: 281). The authors distinguish six types of NE (人名 *renming* "personal names", 人物稱謂 *renwu chengwei* "appellation", 官職 *guanwei* "official title", 爵位/封號 *juewei/fenghao* "aristocratic title", 地名 *diming* "place name", 建築名 *jianzhuming* "building name", and 組織名 *zuzhiming* "organizational name"). The categorization seems somewhat idiosyncratic. Why should "buildings" be distinguished from "places"? Why the proliferation of categories that in themselves are not strictly names, and which naturally results in a large number of "compound" NEs (ibid., pp. 286–287), as personal names are combined with, e.g., an official title? For NLP it would probably be better to focus on the basic categories in the original NER specification. Xiong et al. (2013) also suggest to include a distinction between real and fictional person as part of the NER process. This is of course possible for manual tagging or markup, but it is not clear how such a distinction could be implemented computationally in the context of NER.

There are a number of structural problems that make NER for classical literary texts harder than for the modern, spoken forms of Chinese. In classical Chinese the problem of word segmentation is compounded by the fact that the morphology relies less on multi-character compounds, but rather on the context dependent semantics of single characters. These are not bound by word-classes, and, of course, are not inflected. It is often said that in classical Chinese single characters equal words, though that is only true in a very general sense. Recent research by Chen et al. (2012) have shown that the frequency distribution of words and characters follow different laws[6]. Moreover, the use of multi-character words is often dependent on genre. Between the extremes of Buddhist Hybrid Chinese (佛教混合漢語 *fojiao hunhe hanyu*)[7] with its many transcribed or transliterated multi-character words and the concise diction of the Analects or Mencius there is a wide range of different styles and registers.

Therefore, in both modern and classical Chinese, a character sequence A B C can in principle be read in one of four different ways: AB/C, A/BC, A/B/C or as one single compound ABC. In the relatively limited, and better understood, dictionary of modern

Chinese, however, strong probabilities can be assigned to compounds such as AB, BC or ABC. The more context-dependent semantics of the classical idiom on the other hand make this difficult.

Classical Chinese has not only fewer multi-character words, but on the whole the semantic range of single-character words in classical Chinese is wider than in modern Chinese, allowing for more possible combinations with adjacent characters. Many of the different meanings of each character listed in the 漢語大詞典 *Hanyu Dacidian* "Dictionary of Chinese" are well attested in classical Chinese, but have become obsolete in modern Mandarin. The wide semantic range combined with the absence of word-classes in the Indo-European sense means that characters can appear in various syntactical positions.

Moreover, all characters are, at least in principle, usable in named entities. In literary Chinese there are almost no special characters reserved for names. Although there is a very limited number of family or clan names (姓*xing*), which appear with high frequency, in principle any character can be used in personal names (名*ming*), courtesy names (字 *zi*), pen names (號*hao*), taboo names (諱*hui*), and of course place names. In practice, characters with negative connotations were avoided.

Also, like in English, a name in classical Chinese can be abbreviated or extended in various ways (J. Smith, John R. Smith, John vs. 朱熹 *Zhu Xi*, 朱公 *Zhu gong*, 朱夫子 *Zhu fuzi*, 朱文公 *Zhu wengong*)[8]. For their corpus of Tang poetry Luo et al. 羅鳳珠等 (2013) list 25 different combinations of names and appellatives (ibid., p. 5f). Though comparative studies are not yet available, the relationship between entity and names in classical Chinese sources was probably more complex than in European texs, as it was customary for all notables to have several names, none of which has any phonetic, semantic or grammatic relationship with the personal name. In prose full names are often given only on first occurrence and afterwards abbreviated to a single character. These abbreviations are especially difficult to detect for a rule-based algorithm and beyond the reach of a onomasticon.

Another reason why it is more difficult to solve NER for classical than for modern Chinese is the lack of punctuation. Punctuation is a disambiguating feature that helps some NER algorithms. This is often noticed only once punctuation is absent, e.g. in the case of transcription. The canonical definition of NER remarks: "Transcriptions of speech lack most capitalization and punctuation found in electronic newswire articles; this missing information makes certain decisions regarding proper names more difficult". (Chinchor et al. 1999: 1). Nadeau and Sekine (2007: 8) list punctuation among the "features most often used for the recognition and classification of named entities". Li and Sun (2009: 509–510) have proven that punctuation aids word segmentation in modern Chinese, especially for NER. We have already remarked on the fact that neither modern nor classical Chinese use spaces, but machine learning algorithms for modern Chinese are at least able to latch onto punctuation, which helps in distinguishing word boundaries and identifying probabilistic patterns. Without punctuation what constitutes a "sentence" is difficult to define even for modern languages. All the more so perhaps for classical Chinese, which, outside of ritual declamations, was only ever used as written language. Although the notion of "sentence", as well as punctuation marks, existed in pre-modern China, there never emerged a standardized punctuation system (Harbsmeier 1998: 173–184). The orthographic *de facto* definition of sentences by punctuation has proved useful for computational linguistics in European languages. Sentences in this sense can be used to align multilingual corpora and probabilistic machine learning algorithms are able to latch onto the proto-tagging of punctuation. Names, for instance, might possibly appear closer to the right of a full stop than to its left, as they often function as grammatical subjects. Though modern editions of classical Chinese texts often have some form of punctuation, large corpora such as the 四庫全書 *Siku quanshu* "Imperial Collection in Four Sections" do not. It would make little sense to train an learning algorithm on a punctuated text and then test it on a non-punctuated text.

The difficulties in delimiting "words" and "sentences" in classical Chinese compound two typical NER problems: variation and ambiguity. For the purpose of its use in the Digital Humanities it should be remembered that "disambiguation" in terms of basic NER is only rudimentary. The task for basic NER mainly consists in deciding whether a word is (syntactically) a name of a person, a location or an organization. Applications, however, should not only tell us that a string is a person or a place name, but would ideally help to identify a person or a place unambiguously and tie it to occurrences in other texts. This is beyond the narrow definition of NER and something that can be done only with the help of larger, knowledge-based systems which include some form of authority data. Knowledge-based systems, such as ontologies are often based on different domains, which have their own distinctive use of language.

Buddhist Hybrid Chinese, for instance, the written form of Chinese that was used between 100 and 1100 CE to translate Indian texts, has many irregularities on the semantic, morphologic and syntactic level. While only a small number of new characters were created for the translation of Buddhist texts (e.g. 魔 *mo* "Māra, demon, tempter"), a number of characters have acquired distinctive Buddhist meanings (e.g. 業 *ye* "karma"). The phonetic transcription of Indian terms with the help of Chinese characters contributed to the increased use of two and three character compounds in written Chinese, as a large number of new words were coined by the translators (e.g. 沙門 *shamen* "renunciant") (Liang 梁曉虹 2008: 257). On the syntactic level, Buddhist scriptures at times retain traces of the Indian original and the verb, for instance, can end up after the object, which in classical Chinese happens only rarely and then usually in verse. In order to encode the many rare characters and proof the texts efficiently, one needs a staff of encoders who are familiar with Buddhist Hybrid Chinese. Significant domain knowledge is needed to create digital corpora of ancient Buddhist texts and it is not surprising that all significant Buddhist corpora were created by dedicated institutions, such as the Chinese Buddhist Text Association (CBETA) in Taiwan, the SAT Daizōkyō Text Database in Japan, or the Research Institute of Tripitaka Koreana in Korea.

4 Introduction to the Digital Archive of Buddhist Temple Gazetteers

From 2008 to 2011 researchers at the Dharma Drum Institute of Liberal Arts (formerly Dharma Drum Buddhist College) encoded a corpus of local histories of Buddhist sites, so-called “gazetteers” (志/誌 *zhi*)[9]. This resulted in the “Digital Archive of Chinese Buddhist Temple Gazetteers”, that is described in Bingenheimer (2012) and Hung (2013). The archive is still maintained and developed. One reason to create this corpus was to showcase the use of complex, marked-up editions in the study of Chinese Buddhist history. The digital texts are made available via an online interface, a DVD, and freely downloadable archive files in METS format[10]. This is a rarity, as unfortunately most Chinese corpora are not available under an open access license. The diachronic Sheffield Corpus of Chinese (c. 430,000 characters) is well organized and tagged in great detail, including dates, person and place names, but its online interface offers only a very limited search function, not the data as such. A 2004 version of the corpus is, however, available at the Oxford Text Archive. The pre-modern corpora created at Academia Sinica are only for licensed used, and, moreover, are not tagged for NER (Xiong et al. 2013: 282). The Peking Corpus of Ancient Chinese too provides only a very limited online search interface. The only other freely available corpora of classical Chinese explicitly tagged for NEs is the 佛教傳記文學 *Fojiao zhuanji wenxue* “Gaoseng Zhuan Corpus”[11]. Like the Gazetteer Archive corpus, the Gaoseng Zhuan Corpus is not POS tagged for use in computational linguistics, but relies on detailed TEI/XML markup to identify names and dates in historical research.

Twelve gazetteers from the archive have appeared in a print series, which in the preface is described as “a snapshot … in the development of the digital text”. (Bingenheimer 馬德偉 2013: Vol.1: vii). In the present paper we want to highlight another use of the corpus: as benchmark for research on NER in classical Chinese.

As of today (Feb. 2015) the corpus consists of two sections: The texts of 208 Temple Gazetteers are available without punctuation and only basic structural TEI markup. Besides those, another 15 gazetteers were digitized with added punctuation and sophisticated, manual markup that identifies and disambiguates names and dates and links them to an authority database.

The 208 gazetteers are basically text-only versions with a metadata header and markup indicating distinct texts (<div>) within the gazetteer, their headings (<heading>), authors (<byline>), page- and line-breaks (<pb>, <lb>). The markup differentiates prose (<p>) from table content (<table> a .o.), and verse sections (<lg> a. o.). Furthermore it aligns the page-breaks with the image files that are part of the archive distributables. Metadata on the image files is provided in the METS metadata wrapper, which too is part of the archive. The combined character count of the 208 gazetteers with basic markup is c. 14,000,000[12].

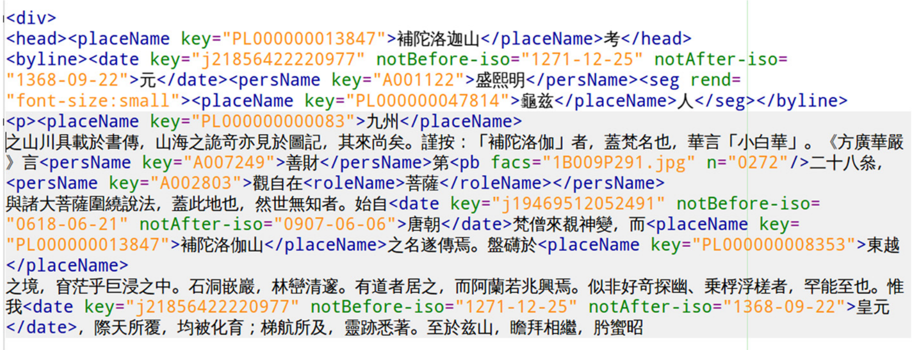
The 15 gazetteers with modern punctuation are built on the same basic tag set, but include additional markup going far beyond it. The research team has painstakingly identified all person and location names (<persName>, <placeName>) as well as all dates (<date>) in the texts. KEY attributes link the names to authority databases (http://authority.ddbc.edu.tw/) that were created at Dharma Drum in the process of this and other projects. All name and date entities are not only identified, but also fully disambiguated. The name “Avalokiteśvara”, for instance, might appear in the sources transcribed or translated in various ways: 阿縛盧枳低濕伐羅 *Afuluzhidishifaluo*, 觀音 *Guanyin*, 觀世音 *Guanshiyin*, 光世音 *Guangshiyin*, 觀自在 *Guanzizai*. These names are all mapped to the same ID (A002803), as are other designations of this Bodhisattva, such as 白衣大士 *Baiyi dashi* “white-robed Mahāsattva/Great Being” or 普陀大士 *Putuo dashi* “Mahāsattva/Great Being of Mount Putuo”.

The process is identical to disambiguation in European languages. The pre-standardized spelling of Shakespeare – Shakespear, Shakspeare, Shakspere a.o.– all refer to the same person, as does “The Bard of Avon”. Beyond merely recognizing NEs as part of POS tagging, which was the original problem of NER, the successful disambiguation of NEs at one point has to include some form of authority data or ontology that can assist the algorithm or, as in our case, the human encoder.

This separation between markup and authority database should be considered best practice as it reduces complexity. Certain distinctions that in NLP are usually implemented in the tagging itself, can be outsourced to the authority databases, for instance whether a person is considered historical or fictional, or a location is a city or a building (cf. Xiong et al. 2013).

The combined character count of the 15 fully marked-up gazetteers is c. 1,600,000[13]. For the amount of NEs and the ratio of NEs to the whole text see Tables 1 and 4 below. Within those 15 gazetteers, rare characters that appear in the woodblock prints were marked-up in TEI in a way that allows regularizing them in different output modes. For the print version, for instance, we were able to map rare variants to their closest equivalent in Unicode, thus avoiding having to font a large number of variants (see Bingenheimer 馬德偉 2013: 凡例 *fanli* “Editorial principles”).

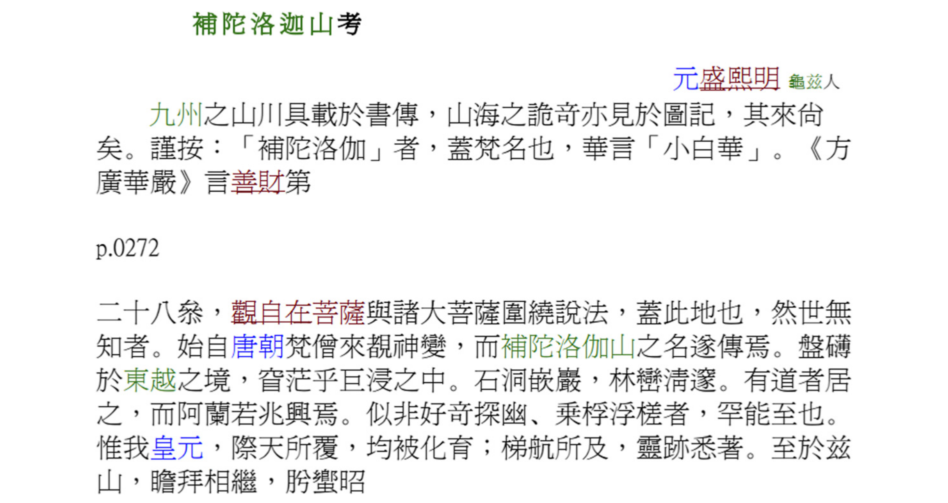
The project was constructed as digital archive, and the same code is used in different output formats. First an example of the XML/TEI master format which in a text editor might appear like this (see Figure 1).



```
<div>
<head><placeName key="PL000000013847">補陀洛迦山</placeName>考</head>
<byline><date key="j21856422220977" notBefore-iso="1271-12-25" notAfter-iso=
"1368-09-22">元</date><persName key="A001122">盛熙明</persName><seg rend=
"font-size:small"><placeName key="PL000000047814">龜兹</placeName>人</seg></byline>
<p><placeName key="PL000000000083">九州</placeName>
之山川具載於書傳，山海之詭奇亦見於圖記，其來尚矣。謹按：「補陀洛伽」者，蓋梵名也，華言「小白華」。《方廣華嚴
》言<persName key="A007249">善財</persName>第<pb facs="1B009P291.jpg" n="0272"/>二十八粂，
<persName key="A002803">觀自在<roleName>菩薩</roleName></persName>
與諸大菩薩圍繞說法，蓋此地也，然世無知者。始自<date key="j19469512052491" notBefore-iso=
"0618-06-21" notAfter-iso="0907-06-06">唐朝</date>梵僧來覩神變，而<placeName key=
"PL000000013847">補陀洛伽山</placeName>之名遂傳焉。盤礴於<placeName key="PL000000008353">東越
</placeName>
之境，窅茫乎巨浸之中。石洞嵌巖，林巒清邃。有道者居之，而阿蘭若兆興焉。似非好奇探幽、乘桴浮槎者，罕能至也。惟
我<date key="j21856422220977" notBefore-iso="1271-12-25" notAfter-iso="1368-09-22">皇元
</date>，際天所覆，均被化育；梯航所及，靈跡悉著。至於玆山，瞻拜相繼，肸蠁昭
```

Figure 1 Code (XML/TEI markup. Data available at: http://buddhistinformatics.ddbc.edu.tw/fosizhi/)

The KEY attributes on the <persName>, <placeName> and <date> elements connect the named entity to the authority databases. The XML/TEI data is transformed to HTML for an online interface that can exploit these connections as links, which allow users to access the information from the database (see Figure 2).



補陀洛迦山考

元盛熙明 龜兹人

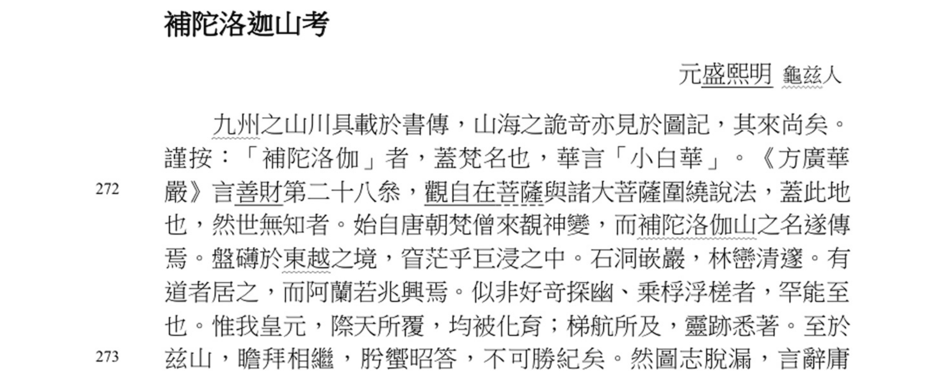
九州之山川具載於書傳，山海之詭奇亦見於圖記，其來尚
矣。謹按：「補陀洛伽」者，蓋梵名也，華言「小白華」。《方
廣華嚴》言善財第

p.0272

二十八叅，觀自在菩薩與諸大菩薩圍繞說法，蓋此地也，然世無
知者。始自唐朝梵僧來覩神變，而補陀洛伽山之名遂傳焉。盤礴
於東越之境，窅茫乎巨浸之中。石洞嵌巖，林巒清邃。有道者居
之，而阿蘭若兆興焉。似非好奇探幽、乘桴浮槎者，罕能至也。
惟我皇元，際天所覆，均被化育；梯航所及，靈跡悉著。至於兹
山，瞻拜相繼，肹蠁昭

Figure 2 Online Interface (at: http://buddhistinformatics.ddbc.edu.tw/fosizhi/ui.html?book=g008)

In the printed edition on-click links are difficult to implement, instead person and place indices are compiled and appended to each volume. Dates are generally mapped to the (proleptic) Gregorian calendar (see Figure 3).



補陀洛迦山考

元盛熙明 龜兹人

九州之山川具載於書傳，山海之詭奇亦見於圖記，其來尚矣。
謹按：「補陀洛伽」者，蓋梵名也，華言「小白華」。《方廣華
272 嚴》言善財第二十八叅，觀自在菩薩與諸大菩薩圍繞說法，蓋此地
也，然世無知者。始自唐朝梵僧來覩神變，而補陀洛伽山之名遂傳
焉。盤礴於東越之境，窅茫乎巨浸之中。石洞嵌巖，林巒清邃。有
道者居之，而阿蘭若兆興焉。似非好奇探幽、乘桴浮槎者，罕能至
也。惟我皇元，際天所覆，均被化育；梯航所及，靈跡悉著。至於
273 兹山，瞻拜相繼，肹蠁昭答，不可勝紀矣。然圖志脫漏，言辭庸

Figure 3 Printed edition (Bingenheimer 馬德偉 2013: Vol. 3, p. 118)

5 Named Entities and NER in the Digital Archive of Buddhist Temple Gazetteers

The separation of semantic information (encoded in TEI) from the presentation layer is the basis for creating different views of the texts (print, online, as audio book etc.). However, markup also allows a more sophisticated, research-oriented analysis of the texts. We can now answer questions such as: How much of the text consists of NEs? How many NEs are there on average per text? Are the averages comparable, or do they vary significantly between texts? How many unique persons and places are there as compared with the number of overall occurrences? Below are the query results and some answers to these questions[14] (see Table 1).

Table 1 Person and Location Entities

Gazetteer (Date)	Person-names (total)	Person-names (unique values)	Person-frequency factor (All names/Unique values)	Location-names (total)	Location-names (unique values)	Location-frequency factor ratio
重修普陀山志 *Chongxiu putuo shan zhi* "Revised Gazetteer of Mount Putuo" (1607)	1164	464	2.51	1584	414	3.83
明州阿育王山志 *Mingzhou ayuwang shan zhi* "Mingzhou Aśoka Temple Gazetteer" (1612)	2741	679	4.04	1956	498	3.93
福建泉州開元寺志 *Fujian quanzhou kaiyuan si zhi* "Fujian Quanzhou Kaiyuan Temple Gazetteer" (1643)	767	320	2.40	548	230	2.38
明州阿育王山續志 *Mingzhou ayuwang shan xuzhi* "Supplement to the Mingzhou Aśoka Temple Gazetteer" (1757)	1785	526	3.39	1104	285	3.87
黃檗山志 *Huangbo shan zhi* "Mount Huangbo Gazetteer" (1824)	1988	582	3.42	1784	385	4.63
慧因高麗華嚴教寺志 *Huiyin gaoli huayanjiao si zhi* "Huiyin Koryo Huanyan School Temple Gazetteer" (1881)	1422	429	3.31	1080	279	3.87
天台山方外志 *Tiantai shan fangwai zhi* "Mount Tiantai Gazetteer" (1894)	6724	1768	3.80	6736	1385	4.86
寒山寺志 *Hanshan si zhi* "Hanshan Temple Gazetteer" (1922)	1751	476	3.68	1022	281	3.64
普陀洛迦新志 *Putuoluojia xin zhi* "New Gazetteer of Mount Putuo" (1924)	5856	2272	2.58	5452	1276	4.27
清涼山志 *Qingliang shan zhi* "Mount Qingliang Gazetteer" (1933)	2981	663	4.50	2634	613	4.30
峨眉山志 *Emei shan zhi* "Mount Emei Gazetteer" (1934)	2696	808	3.34	5078	929	5.47
九華山志 *Jiuhua shan zhi* "Mount Jiuhua Gazetteer" (1938)	3025	916	3.30	4300	821	5.24

Table 1 is about occurrences of person and location names in the gazetteers. The markup allows a similar analysis for date expressions, opening up the prospect of visualizing events along timelines, but for now we will limit ourselves to person and place names. The KEY attribute on <persName> and <placeName> that points to unique authority database entries, allows to remove multiple mentions as well as homonymy. Counting unique values results in the actual number of persons and locations mentioned in the text with great precision (there are a few "unknown" names which resist identification). The range that is described by the "Frequency factor" approximates how many times person or place names are mentioned in the text on average. The higher the ratio the more often a gazetteer repeats its names. The ratio varies slightly between gazetteers: 2.4 to 4.5 with person names; and 2.38 to 5.47 with location names. The lower limits are probably due to the shortness of the 福建泉州開元寺志. The arithmetic average might not be the most useful average measure here. All gazetteers have a "long tail" of many unique occurrences of names, and a small number of names that are mentioned very frequently. More sophisticated math will be able to describe the distribution more precisely.

In a next step, by using the (freely available) Dharma Drum person authority database (http://authority.ddbc.edu.tw/person/) further queries can answer which persons were mentioned most frequently. Below are the twenty most frequent person names in the early 20th century gazetteers of the three most famous Buddhist Mountains: Mount Wutai, Mount Emei and Mount Putuo[15] (see Table 2).

Table 2 Twenty most frequent person names in three gazetteers

普陀洛迦新志(1924)	清涼山志(1933)	峨眉山志 (1934)
Person-names occurances	Person-names occurances	Person-names occurances
觀世音菩薩 *Guanshiyi pusa* (x 167)	文殊菩薩 *Wenshu pusa* (x 365)	普賢菩薩 *Puxian pusa* (x 308)
		(unknown x 87)
性統 *Xingtong* (x 118)	(unknown x 189)	文殊菩薩 *Wenshu pusa* (x 43)
通旭 *Tongxu* (x 99)	張商英 *Zhang Shangyin* (x 68)	福登 *Fudeng* (x 41)
繹堂 *Yitang* (x 73)	鎮澄 *Zhencheng* (x 63)	克勤 *Keqin* (x 37)
法澤 *Faze* (x 53)	德清 *Deqing* (x 51)	廣成子 *Guang Chengzi* (x 33)
化聞 *Huawen* (x 49)	真可 *Zhenke* (x 48)	可聞 *Kewen* (x 29)
(unknown x 47)[a]	福登 *Fudeng* (x 44)	孫思邈 *Sun Simiao* (x 27)
藍理 *Lanli* (x 43)	無著 *Wuzhuo* (x 38)	蔣超 *Jiang Chao* (x 25)
梅福 *Mei Fu* (x 41)	法照 *Fazhao* (x 33)	黃帝 *Huangdi* (x 24)
真融 *Zhenrong* (x 38)	袾宏 *Zhuhong* (x 28)	胡世安 *Hu Shian* (x 23)
愛新覺羅玄燁 *Aixingjueluo Xuanye* (x 38)	普賢菩薩 *Puxian pusa* (x 26)	釋迦牟尼佛 *Shijiamuni fo* (x 23)
裘璉 *Qiu Lian* (x 37)	澄觀 *Chengguan* (x 23)	蘇軾 *Su Shi* (x 20)
印光大師 *Yinguang dashi* (x 35)	義存 *Yizun* (x 22)	通天 *Tongtian* (x 20)
善財童子 *Shancai tongzi* (x 35)	道義 *Daoyi* (x 22)	樵陽子 *Qiao Yangzi* (x 20)
開如 *Kairu* (x 35)	阿彌陀佛 *Amituo* fo (x 21)	呂洞賓 *Lü Dongbin* (x 19)
了餘 *Liaoyu* (x 35)	法本 *Faben* (x 20)	慧通 Huitong (x 18)
立山 *Lishan* (x 34)	道開 *Daokai* (x 19)	普眼菩薩 *Puyan pusa* (x 18)
陶鏞 *Tao Yong* (x 33)	佛陀波利 *Fotuoboli* (x 18)	觀世音菩薩 *Guanshiyin pusa* (x 17)
清了 *Qingliao* (x 32)	攝摩騰 *Shemoteng* (x 18)	性一 *Xingyi* (x 17)
真可 *Zhenke* (x 31)	阿育王 *Ayu wang* (x 17)	

[a]The entry "unknown" is used for all names that the encoders were not able to identify.

Studying these lists reveals who the gazetteer compiler deemed important for the site. Among the most frequent names are Bodhisattvas, famous abbots, monks and laypersons associated with the site. An immediate, if trivial, result is that one can easily spot which Bodhisattva is associated with which site. On a second glance, somewhat less obvious, it appears that the three sites differ in the prominence they give to *other* Bodhisattvas and Buddhas. Thus in the 峨眉山志, besides 普賢菩薩 *Puxian pusa* "Samantabhadra", there is also 釋迦牟尼佛 *Shijiamoni fo* "Shakyamuni", 文殊菩薩 *Wenshu pusa* "Mañjuśrī", 觀世音菩薩 *Guanshiyin pusa* "Avalokiteśvara", and 普眼菩薩 *Puyan pusa* "Samantanetra", among the twenty most frequently mentioned names. The

清涼山志, besides Mañjuśrī, frequently mentions 普賢菩薩 *Puxian pusa* “Samantabhadra” and 阿彌陀佛 *Amito fo* “Amitabha Buddha”. In contrast, the list for the 普陀洛迦新志 includes no other savior figures apart from Avalokiteśvara. It appears that Mount Putuo is more exclusively associated with Avalokiteśvara than the other two sites are with Mañjuśrī and Samantabhadra respectively.

That Mount Putuo has been more focused on Avalokiteśvara than other Buddhist mountains on “their” Bodhisattva, is corroborated by a look at the temple architecture and other features of the site. Guanyin is the central image in the main hall of the major temple at Mount Putuo and the other two major temples too are tied to the Guanyin cult. Both Mount Wutai and Mount Emei accommodate a larger and more diverse number of temples, and are iconographically less committed to a single savior figure.

Another question that can be asked is regarding the connection of the sites with literary figures. In the case of Mount Wutai the name of 張商英 *Zhang Shangying* “Zhang Shangying” is firmly associated with the site, because of his travelogue (Gimello 1992). In the 峨眉山志, 蘇軾 *Su Shi* “Su Shi” appears prominently, but not 范成大 *Fan Chengda* “Fan Chengda”, who wrote an important travelogue (Fan appears only on position 31) (Hargett 2006). Among the “Top 20” of the Putuo Gazetteer, on the other hand, there is no famous literary figure.

In cases where a site has gazetteers of different periods we can attempt a diachronic view and compare how the “landscape of names” has changed over the centuries (see Table 3).

Table 3 Diachronic view of the 20 most frequent person names in two gazetteers on Mount Putuo

重修普陀山志 (1607)	普陀洛迦新志 (1924)
Person-names occurrences	Person-names occurrences
觀世音菩薩 *Guanshiyin pusa* (x 86)	觀世音菩薩 *Guanshiyin pusa* (x 167)
張隨 *Zhang Sui* (x 39)	性統 *Xingtong* (x 118)
屠隆 *Tu Long* (x 24)	通旭 *Tongxu* (x 99)
(unknown x 21)	繹堂 *Yitang* (x 73)
梅福 *Mei Fu* (x 19)	法澤 *Faze* (x 53)
善財童子 *Shancai tongzi* (x 17)	化聞 *Huawen* (x 49)
真融 *Zhenrong* (x 16)	(unknown x 47)
周應賓 *Zhou Yingbin* (x 16)	藍理 *Lan Li* (x 43)
性能 *Xingneng* (x 15)	梅福 *Mei Fu* (x 41)
如曜 *Ruyao* (x 14)	真融 *Zhenrong* (x 38)
龍德孚 *Long Defu* (x 13)	愛新覺羅玄燁 *Aixinjueluo Xuanye* (x 38)
李綵鳳 *Li Caifeng* (x 12)	裘璉 *Qiu Lian* (x 37)
龍女 *Longnü* (x 12)	印光大師 *Yinguang dashi* (x 35)
如迥 *Rujiong* (x 11)	善財童子 *Shancai tongzi* (x 35)
文殊菩薩 *Wenshu pusa* (x 11)	開如 *Kairu* (x 35)
釋迦牟尼佛 *Shijiamuni fo* (x 11)	了餘 *Liaoyu* (x 35)
盛熙明 *Sheng Ximing* (x 10)	立山 *Lishan* (x 34)
孔丘 *Kong Qiu* (x 10)	陶鏞 *Tao Yong* (x 33)
清了 *Qingliao* (x 9)	清了 *Qingliao* (x 32)
慧鍔 *Hui’e* (x 9)	真可 *Zhenke* (x 31)

Apart from legendary and religious figures (觀音 *Guanyin* "Avalokiteśvara", 善財 *Shancai* "Sudhana", 梅福 *Mei Fu* "Mei Fu") only two names have made it into the list both in the early 17th and the 20th century gazetteer. Although neither 真融 *zhenrong* "Zhenrong" (1524–1592) nor 真歇清了 *Zhenxie Qingliao* "Zhenxie Qingliao" (1088–1151) are household names, the above highlights them as important for the cultural memory of Mount Putuo and might guide further research in their direction.

Another question it now becomes possible to ask is: How much of the text consists of named entities and dates? This is important for comparative studies between corpora or to flag single gazetteers that have "eccentric" NE patterns. For our twelve gazetteers the tally is shown in Table 4.

Table 4 Named entities and dates: percentage of text

Gazetteer (Date)	Text length (UTF-8 characters)	Person-names of text length	Location-names of text length	Dates of text length	Names and dates of text length
重修普陀山志 (1607)	70885	03.747 %	05.465 %	01.684 %	10.897 %
明州阿育王山志 (1612)	116485	05.025 %	04.066 %	01.522 %	10.613 %
福建泉州開元寺志 (1643)	31832	05.111 %	03.993 %	03.845 %	12.949 %
明州阿育王山續志 (1757)	54770	07.387 %	04.858 %	02.061 %	14.307 %
黃檗山志 (1824)	102305	04.105 %	04.055 %	01.747 %	09.907 %
慧因高麗華嚴教寺志 (1881)	54645	05.772 %	04.985 %	01.760 %	12.517 %
天台山方外志 (1894)	301992	04.559 %	05.203 %	01.773 %	11.535 %
寒山寺志 (1922)	58231	06.187 %	04.099 %	01.918 %	12.205 %
普陀洛迦新志 (1924)	259705	04.637 %	05.489 %	02.800 %	12.926 %
清涼山志 (1933)	137460	04.346 %	04.537 %	01.452 %	10.335 %
峨眉山志 (1934)	161778	03.645 %	08.262 %	01.253 %	13.159 %
九華山志 (1938)	153452	04.235 %	06.914 %	02.158 %	13.307 %

The result shows that temple gazetteers in the last three centuries have a similar text/name-date ratio. The overall range of 9.9 % to 14.3 % can be narrowed significantly by disregarding two outliers, the 黃檗山志 (g086) and the 明州阿育王山續志 (g011). The 黃檗山志 is focused on the figure of Yinyuan and his disciples and therefore mentions fewer different names, using pronouns and demonstratives, which lowers the character count for proper names. The 明州阿育王山續志 consists almost entirely of poetry. Most of the short poems are associated with a place and an author, which might account for the relative high amount of names and dates. Without these two the range is between 10.3 % and 13.3 %, merely 3 percent. Is this typical for multi-genre compilations of classical Chinese texts? We do not know, because the gazetteer corpus is so far the only freely accessible corpus of classical Chinese that is tagged for NEs. How similar are Buddhist temple gazetteers in this, for instance, to the corpus of letters and notebooks from the Song and Yuan dynasties[16]? Without a tagged corpus of notebooks this is difficult to answer, but it would be interesting to know if there is a genre specific "name density", which would allow for automated genre detection, or help to flag "eccentric compilations", which contain texts with unusual high or low amounts of names and dates.

Could the Dharma Drum gazetteer corpus be used as a training corpus for NER on Song-Yuan notebooks and letters? In principle yes, but with some caveats. Gazetteers generally are compilations of text from very different genres. This makes them interesting historical sources, but for a training corpus for 筆記 *biji* literature one must consider some adjustments. All verse passages (<l>s within <lg>s), for instance, should be removed from the gazetteer corpus, if it were to serve as training data for a corpus of *biji* literature (which consists almost exclusively of prose). Another feature that has to be removed from or disregarded in the gazetteer corpus is punctuation, as machine learning algorithms would latch on to that for feature recognition (if not specifically instructed not to) as has been shown for modern Chinese (Rao and Xun 2012: 18–19). Nevertheless the gazetteer corpus could certainly serve as next-of-kin for a corpus of Song-Yuan notebooks, or the corpus of official dynastic histories (正史 *zhengshi*) and allow benchmarking and aid digital analysis.

For the reasons outlined in Section 2 above there is little reason to be overly optimistic about NER for classical Chinese in the near future. What does a "brute force attack", such as using a dedicated dictionary to parse for names, tell us about the material? For the marked-up part of the gazetteer corpus we can attempt such an approach with an ideal dictionary, i.e. a authority database of names that was created in conjunction with the markup of the corpus. By using the Dharma Drum person authority database as dictionary we can reduce the number of *false negatives* to a minimum, because, apart from a few exceptions, all names in the corpus appear in the dictionary[17]. That way we can concentrate on the problematic recall rate, the high number of the *false positives*, that parsing classical Chinese is likely to yield. To clarify: The dictionary lookup identifies strings in the corpus. The conditions are ideal because all names are in the dictionary (preventing false negatives) and the corpus contains punctuation (preventing some false positives, because names do not cross punctuation boundaries). The database records a number of alternative names for every person name.

Because of the markup we have a clear benchmark in the sum of the person names occurrences (s. Table 1). According to the markup the twelve gazetteers contain 32,900 names, which denote 7773 unique persons. The figure of unique names is lower than the sum of all unique names in Table 1, because the overlap of names between gazetteers has to be accounted for. Before parsing the corpus with the dictionary, we delete all single character names from the person authority dataset (Version 2013)[18]. This reduces the 52,535 names of the dictionary to 51,620, but does greatly reduce the number of false positives. A pass of the entire dictionary (including the single characters) over the corpus gives 821,210 name occurrences. After deleting the single character names the yield is only 133,388 occurrences, which is still about four times the number derived from markup (32,900). A false positive ratio of c.75 % is not acceptable even for modest standards of strict NER, which does not seek to disambiguate persons from each other. One could perhaps imagine a tool that helps with the markup by preprocessing the texts and then leaving to encoders to disambiguate persons manually, i.e. deciding which flagged names are actually person names and which not (e.g. the string 白雲 *baiyun* "white cloud" appears 301 times in the corpus, but only 16 times within person names). Nevertheless in a real world project the dictionary is not, like in our example, exhaustive and unflagged text too must be checked for occurrences (false negatives).

Thus even under optimal conditions – using a custom dictionary that was created during the markup of the very texts to be analyzed – person names in classical Chinese cannot be reliably identified for statistical analysis through dictionary look-up alone. The very factors that make classical Chinese resistant to word segmentation, NER and disambiguation, prevent even a dictionary that has been created for a corpus to reliably perform NER on the very same corpus.

Advanced probabilistic, machine-learning methods, that (to a degree) can be adapted to an environment without word boundaries are obviously the way forward here. These algorithms will need training data, however, such as the Gazetteer Archive. Semantic ambiguity (person or place? Person A or person B?) beyond NE recognition as part of POS tagging must be addressed with an array of authority lookup tools, such the Buddhist Studies Authority Databases.

6 Conclusion

On the whole, sophisticated parsing of classical Chinese is still in its early stages. Judging from the state of NER for modern Chinese, and taking into account the linguistic characteristics of the classical idiom, it is of course no surprise that that even a custom-made dictionary is not able to parse (even a punctuated) text well. Even using more sophisticated approaches it seems unlikely that machines will be able to successfully handle NER for classical Chinese in the near future. For all that has been written about NER in classical Chinese, and the various proposals for segmentation, I have found not one case where NER has been successfully applied to a classical text (i.e. with a reasonably high F-score). On the other hand, however, we have seen that marked-up corpora can answer research questions in the Humanities that could only be guessed at before. The production, querying and transformation of marked-up corpora is well within the purview of DH computing as it relies on algorithms that are considerably less complex, than those needed for NER on unmarked texts. What we can hope for with regard to NER is that through advancements in machine-learning based systems, computers will be able to flag the most likely candidates for names and dates resulting in semi-supervised markup systems[19]. This would be useful, because for many tasks, high accuracy is not all that important. Topic discovery, for instance, may very well be possible with low recall rates as long as precision is reasonably high. Nevertheless, it makes sense for individual researchers and institutions to invest in markup. Although expensive, marked-up corpora are reusable and through collaboration and shared standards it is possible to aggregate them into larger datasets that function as improved editions as well as provide for reuse and repurposing by other researchers. Crucially, future NER will need annotated corpora as training data. Here the Digital Archive of Temple Gazetteers, as well as the Corpus of Biographies of Eminent Monks produced by the same team, might be able to help. The production of better and more detailed marked-up corpora is field where Humanities and NLP can cooperate. Such corpora support both the traditional tasks of scholarship in the Humanities such as editing, translating and historical research, as well as emerging forms of scholarship such as data mining, distant reading and visualization.

7 Endnotes

[1] This paper was first presented at the conference *Letters and Notebooks as Sources for Elite Communication in Chinese History, 900–1300* held at Oxford University in January 2014, I thank the organizer Hilde De Weerdt and the participants for the valuable feedback.

[2] Another prominent digital corpus, the famous *Corpus Thomisticum*, was started even earlier, but used for linguistic analysis only later.

[3] The term "gazetteer" is a terminological problem for this paper. In NER lists of person and especially place names that are used to identify and disambiguate NEs are called "gazetteers". The corpus I describe here unfortunately uses the word "gazetteer" in a different sense. In sinological usage the word is used to translate 志 (or 誌) *zhi*, a local history comprising different genres of texts about a location.

[4] GB13715 seems to have been superseded to a degree by Yu et al. (2003). In Taiwan a national standard 中文分詞處理原則 *Zhongwe fenci chuli yuance* 'Segmentation principle for Chinese language processing' (CNS14366) (1999)) was developed by Huang et al. (1997).

[5] Huang, Peng et al. (2002) believe the opposite and suggest to equate character and word for POS tagging. This is obviously the result of a very limited view of classical Chinese and explains assumptions such as "Fortunately however, most of the important Classical Chinese documents have already been manually punctuated in the 20th century," (p.117). For a more sophisticated approach see Chen, Gang et al. (2012).

[6] The authors also propose that bi-noms and tri-noms are used increasingly in written Chinese since the Tang dynasty. This increase in word-length might, however, be due to genre rather than language change. The study compares verse of the Tang, Song, and Yuan with prose texts of the Ming, Qing, and the modern era. Comparing verse and prose an increase in word length is perhaps not surprising since verse (especially in the Tang) tends to be more concise. To answer the question if there was an increase in compound words we need to compare medieval Tang/Song/Yuan prose with Early Mandarin prose.

[7] A term sometimes used to emphasize the hybrid nature of Buddhist Chinese, as Indian-Chinese and Literary-Vernacular texts (Mair 1994: 712 *et pass.*; Zhu 朱慶之 2001, Teng 2014).

[8] See Xiong, Xu et al. (2014) on attempts to parse honorifics in classical Chinese.

[9] See footnote 5 above for the use of the term "gazetteer" in this paper. The project was funded by the Chung-hwa Institute of Buddhist Studies and proposed and supervised by the author.

[10] Available at http://buddhistinformatics.ddbc.edu.tw/fosizhi/, (Feb. 2015).

[11] Available at http://buddhistinformatics.ddbc.edu.tw/biographies/gis, (Feb. 2015).

[12] The exact number (14,043,777) is the sum of the approximate character counts for each downloadable archive provided on the website under "Full-text and Image Archives" (http://buddhistinformatics.ddbc.edu.tw/fosizhi, Feb 2015). The count of 251 archives in the section header includes references to "doubles," i.e. identical gazetteers from other collections that are already digitized. For the exact relationship between collections see Bingenheimer (2012).

[13] The exact number (1,603,792) is the sum of the approximate character counts for each downloadable archive provided on the website under "Marked-up Full Text and Archives" (http://buddhistinformatics.ddbc.edu.tw/fosizhi, Feb 2015). It includes neither markup nor punctuation marks.

[14] The data used for the queries are the TEI files for the 12 gazetteers that were published in print (Bingenheimer 2013). The files are available under a CC-license as part

of the archives at http://buddhistinformatics.ddbc.edu.tw/fosizhi/. I have used the archives as available online in November 2013. The results here and below are the output of simple queries in XSLT and Python.

[15] These queries necessitate, next to the data used in Table 1, the Dharma Drum Person Authority dataset distributed at http://authority.ddbc.edu.tw/docs/open_content/download.php. For the analysis here and below I have used Version July 2013.

[16] Such a corpus is part of the "Communication and Empires" project (http://chinese-empires.eu/), see De Weerdt (forthcoming 2015).

[17] The bulk of the Dharma Drum person and place authorities were created during this and other markup projects between 2007 and 2012. We have to limit ourselves here to person names because the Dharma Drum place name authority distributable is not complete due to copyright constraints. Dharma Drum can only distribute the entries added for Dharma Drum projects (especially a large number of temple names and Buddhist sights) while the online database look-up and API uses a larger set that includes geo-references for historical places names created at Academia Sinica, Taipei.

[18] A design mistake in the database has it listing single character names, such as 金 *Jin*, 文 *Wen* or 海 *Hai* as alternative names. A single character can be marked-up as a name in a text, but should not be included in the person name authority as such, obviously 王 *Wang* is a 'alternative' for 王安石 *Wang Anshi* but it is counterproductive to include these in an authority database. Later versions will correct this. Further data clean-up necessitates the deletion of spaces, newlines etc.

[19] One such system is currently under development (http://chinese-empires.eu/analysis/notebooks/) at Leiden University. MARKUS is "designed to automate the markup of different kinds of named entities" with a special focus on classical Chinese.

Received: 10 December 2014 Accepted: 5 June 2015
Published online: 17 November 2015

References

Bingenheimer, Marcus. 2012. Bibliographical notes on Buddhist Temple Gazetteers, their prefaces and their relationship to the Buddhist canon. Chung-hwa Buddhist Journal 25: 49–84.

Bingenheimer, Marcus 馬德偉 (ed.). 2013. The Zhonghua collection of Buddhist Temple Gazetteers 中華佛寺志叢書. Vol.1-12. Taipei: Shin Wen Feng.

Chen, Qinghua, Jinzhong Guo, and Yufan Liu. 2012. A statistical study on Chinese word and character usage in literatures from the Tang dynasty to the present. Journal of Quantitative Linguistics 19(3): 232–248.

Chinchor, Nancy, Erica Brown, Lisa Ferro, and Patty Robinson. 1999. 1999 named entity recognition task definition – Version 1.4 whitepaper. The MITRE Corporation and SAIC, 1999.

De Weerdt, Hilde. In print 2015. Information, territory, and networks: The crisis and maintenance of empire in Song China. Harvard University Asia Center.

Gimello, Robert M. 1992. Chang Shang-ying on Wu-t'ai Shan. In Pilgrims and sacred sites in China, ed. Susan Naquin and Chün-fang Yü, 89–149. Berkeley: University of California Press.

Harbsmeier, Christoph. 1998. Logic and language, Vol. 7, Part 1, Science and civilization in China, ed. Joseph Needham. Cambridge: Cambridge University Press.

Hargett, James M. 2006. Stairway to heaven – A journey to the summit of Mount Emei. Albany: State University of New York Press.

Hilbert, Martin, and Priscilla López. 2011. The world's technological capacity to store, communicate, and compute information. Science 332: 60–65.

Huang, Chu-Ren, and Nianwen Xue. 2012. Words without boundaries: Computational approaches to Chinese word segmentation. Language and Linguistics Compass 6(8): 494–505.

Huang, Chu-Ren, Keh-jiann Chen, Feng-yi Chen, and Li-Li Chang. 1997. Segmentation standard for Chinese natural language processing. Computational Linguistics and Chinese Language Processing 2(2): 47–62.

Huang, Liang, Yinian Peng, Huan Wang, and Wu Zhenyu. 2002. Statistical part-of-speech tagging for Classical Chinese. In Text, speech and dialogue: 5th International Conference TSD 2002, LNAI 2448, ed. Petr Sojka, Ivan Kopeček, and Karel Pala, 115–122. Berlin: Springer.

Hung, Jen-jou 洪振洲. 2013. The establishment of a digitization system for Chinese Buddhist Temple Gazetteers 中國佛教寺廟志數位化系統之建置. Dharma Drum Journal of Buddhist Studies 法鼓佛教學報 12: 145–187.

Kucera, Henry, and W Nelson Francis. 1967. Computational analysis of present day American English. Providence: Brown University Press.

Levow, Gina-Anne. 2006. The Third International Chinese Language Processing Bakeoff: Word segmentation and named entity recognition. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, ed. The Association for Computational Linguistics, 108-117. Burwood, Australia: BPA Digital.

Li, Zhongguo, and Maosung Sun. 2009. Punctuation as implicit annotations for Chinese word segmentation. Computational Linguistics 35(4): 505–512.

Liang, Xiaohong 梁曉虹. 2008. Studies in the phonetics and semantics of the language in Buddhist sutras and the binomization of Chinese 佛經音義與漢語雙音化研究. In Buddhism and research in the history of Chinese – with a special focus on Japanese sources 佛教與漢語史研究：以日本資料爲中心, 248–275. Shanghai: Shanghai guiji.

Liu, Yuan, Qiang Tan, and Xukun Shen. 1994. Segmentation standard for modern Chinese information processing and automatic segmentation methodology (GB13715). Beijing: Tsinghua University Press.

Liu, Ping-Ping, Wei-jun Li, Nan Lin, and Xing-Shan Li. 2013. Do Chinese readers follow the national standard rules for word segmentation during reading?. PLoS ONE 8(2): e55440. doi:10.1371/journal.pone.0055440.

Luo, Fengzhu, Xiaoyu Qiu, and Yixian Lin 羅鳳珠, 邱筱榆, 林宜嫺. 2013. Word separation guidelines and named entities characteristics of Tang Song Dynasty poetry 唐宋詩詞分詞規則及命名實體特徵. Paper presented at the 14th Chinese Lexical Semantics Workshop (CLSW 2013), held on May 10–13 at Zhengzhou University, Henan, China.

Mair, Victor H. 1994. Buddhism and the rise of the written vernacular in East Asia: The making of national languages. The Journal of Asian Studies 53(3): 707–751.

Nadeau, David, and Satoshi Sekine. 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1): 3–26.

Norman, Jerry. 1988. Chinese. Cambridge: Cambridge University Press.

Rainie, Lee, and Maeve Duggan. 2012. E-book reading jumps; print book reading declines. Pew Research Center's Internet and American Life Project Report. Available at http://libraries.pewinternet.org/2012/12/27/e-book-reading-jumps-print-book-reading-declines/. Accessed December 2013.

Rao, Gaoqi, and Endong Xun. 2012. Word boundary information and Chinese word segmentation. International Journal on Asian Language Processing 22(1): 15–32.

Saenger, Paul. 2000. Space between words: The origins of silent reading. Paolo Alto: Stanford University Press.

Sproat, Richard, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. Computational Linguistics 22(3): 377–404.

Teng, Wei-Jen. 2014. Medieval Chinese Buddhist exegesis and Chinese grammatical studies. Taiwan Journal of Buddhist Studies 臺大佛學研究 28: 105–142.

Wei, Pei-chuan, P.M. Thompson, Cheng-hui Liu, Chu-ren Huang, and Chaofen Sun 魏培泉, 譚樸森, 劉承慧, 黄居仁, 孫朝奮. 1997. Historical corpora for synchronic and diachronic linguistics studies 建構一個以共時與歷時語言研究為導向的歷史語料庫. Computational Linguistics and Chinese Language Processing 中文計算語言學期刊 2(1): 131–145.

Xia, Fei. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0). PA: University of Pennsylvania. Available at http://www.cis.upenn.edu/~chinese/segguide.3rd.ch.pdf. Accessed June 2014.

Xiong, Dan, Qin Lu, Fengju Lo, Dingxu Shi, Tin-shing Chiu, and Wanyi Li. 2013. Specification for segmentation and named entity annotation of Chinese classics in the Ming and Qing dynasties. In Chinese lexical semantics - 13th Workshop, CLSW 2012, ed. Donghong Ji and Guozheng Xiao, 280–293. Berlin, Heidelberg: Springer.

Xiong, Dan, Jian Xu, Qin Lu, and Fengju Lo. 2014. Recognition and extraction of honorifics in Chinese diachronic corpora. Springer Lecture Notes on Chinese Lexical Semantics, Lecture Notes in Computer Science 8922: 305–316.

Yu, Shiwen, Huiming Duan, Xuefeng Zhu, Bin Swen, and Bao-Bao Chang. 2003. Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation. Journal of Chinese Language and Computing 13(2): 121–158.

Zhu, Qingzhi 朱慶之. 2001. A preliminary discussion of Buddhist Hybrid Chinese 佛教混合漢語初論. Linguistic Analects 語言學論叢 24: 1–33.