Research Article Open Access

John Lee*, Tak-sum Wong

Conversational Network in the Chinese Buddhist Canon

DOI 10.1515/opli-2016-0022
Received February 29, 2016; accepted September 25, 2016

Abstract: This article describes a method to analyze characters in a literary text by considering their verbal interactions. This method exploits techniques from computational linguistics to extract all direct speech from a treebank, and to build a conversational network that visualizes the speakers, the listeners and their degree of interaction. We apply this method to create and visualize a conversational network for the Chinese Buddhist Canon. We analyze the protagonists and their interlocutors, and report statistics on their number of utterances and types of listeners, how their speech was reported, and subcommunities in the network.

Keywords: treebank; direct speeech; Chinese Buddhist Canon; conversational network

1 Introduction

As more literary texts become available in digital form (Crane, 2006), there is growing interest in performing “distant reading” (Moretti, 1999) on these texts. Distant reading complements traditional literary analysis methods with abstract, quantitative representations of a text. These representations range from word usage statistics, which can shed light on authorship questions (e.g., Holmes, 1994; Hung et al., 2010); the social network in a novel, which can capture relationships among the characters and aspects of the plot (e.g., Agarwal et al., 2012); and statistics on joint stage appearances of characters, which can compare the structures of different Shakespearean plays (e.g., Moretti, 2011). These methods are often amenable to automatic natural language processing techniques, which enable them to tackle large collections of texts for whom manual analysis is infeasible.

This article describes a method to analyze relations between characters in a large corpus in terms of their verbal interactions. This method exploits techniques from computational linguistics to automatically extract all direct speech from a treebank, and then builds a “conversational network”. This network is a graph whose nodes represent characters, and whose edges indicate dialog interactions between characters (Elson et al., 2010). It visualizes the prominent speakers, listeners and their degree of interaction.

Far more than just a visualization tool, a conversational network also facilitates a number of possible research directions. One direction is to analyze the speech content of individual characters, for example to investigate distinctive lexical bundles and discourse features (e.g., Csomay, 2013; Sealey, 2010), or to compute character sketches based on keyword usage (e.g., Baker et al., 2013). Another direction is to analyze the network structure to illuminate relations between characters, for example, who is speaking to whom, how often, and how the speech is reported. We pursue the latter direction in this article, and show

Article note: This paper belongs to the special issue on Treebanking and Ancient Languages, ed. by Giuseppe G.A. Celano and Gregory Crane.

***Corresponding author: John Lee,** City University of Hong Kong, Hong Kong, E-mail: jsylee@cityu.edu.hk
Tak-sum Wong, City University of Hong Kong

how a conversational network can reveal the protagonists, the characters with whom they spoke, and the communities knit by frequent conversations.

To illustrate the potential of this method, we apply it to create a conversational network from the Chinese Buddhist Canon — a large corpus that would be difficult for any individual scholar to absorb and manually analyze. Our study constitutes the first quantitative analysis on the direct speech in this corpus. The rest of this article is organized as follows. Section 2 outlines related work. Section 3 summarizes the treebank used in this study. Section 4 describes steps for constructing the conversational network, and Section 5 evaluates its accuracy. Section 6 visualizes and analyzes the network, followed by conclusions in Section 7.

2 Background

A quantitative analysis on social relationships requires an objective criterion to determine whether two characters are “related”. This criterion varied widely in previous research: Two characters may be considered “related” when both appear within a window of *n* words (Oelke et al., 2013); when they appear in the same scene in a novel (Knuth, 1993); when they are recorded to have been present at the same location (Bingenheimer et al., 2011); or when they both participate in an event (Doddington et al., 2004; Agarwal et al., 2010).

In a conversational network, two characters are considered related if they have verbally interacted. Most attempts in constructing conversational networks have targeted structured texts, such as internet relay chat (Mutton, 2004) and e-mail messages (Diesner et al., 2005). These texts are called “structured” because the speakers (senders) and listeners (receivers) are clearly defined. Other genres that are “structured” in this sense include plays, where two characters are related if one is speaking and the other is also on stage (Stiller et al., 2003; Rydberg-Cox, 2011). The networks of *Hamlet* and *Macbeth*, for example, have been comparatively analyzed to shed light on the different community structures in the two plays (Moretti, 2011).

In unstructured texts such as novels, dialogs and their participants are not precisely defined and must be manually annotated, as was done for *Alice in Wonderland* (Agarwal et al., 2012) and parts of *The Story of the Stone* (Moretti, 2011). These annotations have enabled studies on the protagonists and the characters associated with them, and the perspective holder. To tackle larger corpora, automatic annotation are required. Mahlberg and Smith (2012) designed an automatic procedure to extract direct speech in the works of Dickens, and analyzed his use of the suspended quotation. Elson et al. (2010) developed an automatic method of dialog extraction and speaker attribution, and applied it on 60 novels to investigate correlations between the number of characters, the amount of dialog interactions and the novel setting. We adopt a similar methodology in this study and give details about our algorithm in Section 4.

3 Treebank

We apply our proposed method on the Chinese Buddhist Canon[1]. Written in medieval Chinese, the *Tripiṭaka Koreana* is an edition of the Chinese Buddhist Canon derived from the most complete set of available printing blocks, those currently stored at Haein Monastery in Korea (Lancaster and Park, 1979). It has a total of about 50 million Chinese characters. We use the digital version provided by Lancaster (2010).

Our network extraction method requires syntactic information. Since off-the-shelf Chinese syntactic parsers are intended for modern Chinese, we trained a part-of-speech tagger and parser using a dependency treebank of Chinese Buddhist texts (Lee and Kong, 2016). Specifically, we built a word segmenter and part-of-speech tagger with CRF^{++} (Lafferty et al., 2001) using the feature set proposed by Zhao et al. (2007). We then trained a Minimum-Spanning Tree parser for dependency parsing (McDonald et al., 2006). Since the *Tripiṭaka Koreana* has no punctuation, we inserted punctuation from another digital edition of the Chinese Buddhist Canon, the *Taishō Revised Edition*, provided by the Chinese Buddhist Electronic Text Association

[1] We use the Chinese Canon since many Buddhist texts, especially Mahāyāna ones, survive only in their Chinese translations, and have not been translated into English.

(CBETA)[2]. Figure 1 shows an example Chinese parse tree, which follows conventions for part-of-speech tags and dependency labels of the Penn Chinese Treebank (Xue et al., 2005) and Stanford dependencies for Modern Chinese (Chang et al., 2009).

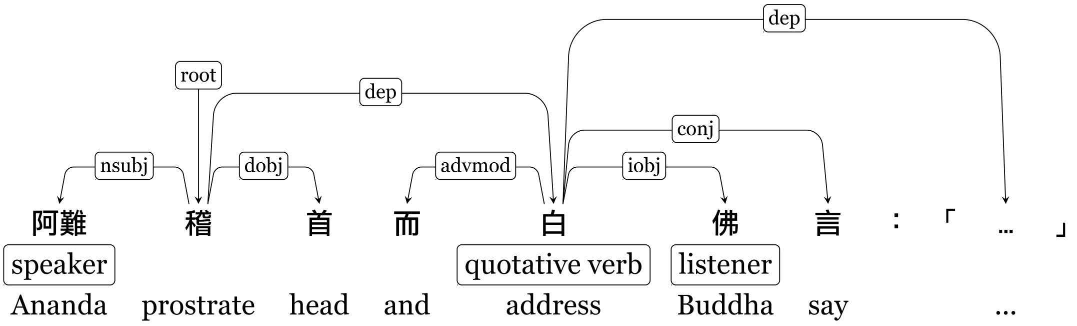

Figure 1. An automatically derived parse tree for a sentence with direct speech, from the Chinese Buddhist Canon.

For the purpose of locating direct speech in the text, the quotation mark is a useful marker. Unfortunately, about two thirds of the CBETA corpus adopts the old Chinese punctuation style, which does not mark direct speech with any punctuation. For this portion of the corpus, we relied on the presence of quotative verbs to detect direct speech (cf. Section 4.1).

4 Network extraction

The vast amount of text in our corpus makes manual analysis unfeasible. Drawing on techniques from computational linguistics, we designed an automatic algorithm that extracts all quoted utterances from the corpora, and attributes to each utterance a speaker and a listener (Table 1). Similar to Elson et al. (2010), we focused on direct speech rather than mixed quotations or indirect speech (Pareti et al., 2013).

Table 1. Automatic utterance retrieval, and attribution of speaker and listener.

Example passage	Utterance	Speaker	Listener
爾時，父王問彼侍者：「太子出遊，歡樂不耶？」答曰：「不樂」。又問其故。問曰：「道逢老人。是已不樂。」 At that time, his father, the king, asked the attendant: “The prince went on an outing. Is he happy?” He replied, “He is not.” The king then asked why. He replied, “On the road he met an old man. It is for this reason that he is unhappy.” *Dīrghāgama* (K647)[3]	太子出遊，歡樂不耶？ The prince went on an outing. Is he happy?	父王 king	侍者 attendant
	不樂 He is not.	侍者 attendant	父王 king
	道逢老人。是已不樂。 On the road he met an old man. It is for this reason that he is unhappy.	父王 king	侍者 attendant

A simple algorithm is to search for quotation markers that are found close to verbs of communication, and scan for personal names against a list of known characters (Pouliquen et al., 2007). The accuracy can be further improved by considering topical similarity (Celikyilmaz et al., 2010). Using a supervised method that identifies trigrams of character mentions, quotative verbs and quotations, Elson et al. (2010) achieved 96% precision and 57% recall in extracting conversational networks from novels. In an approach similar to

[2] Although this version was derived from the same set of printing blocks as the *Tripiṭaka Koreana*, it does not represent the whole of the text glyphs found in the blocks. When the *Taishō Revised Edition* was produced in the 19th century, only 10,000 characters were available to the publishers and thus many substitutions of similar characters had to be made. In contrast, the digital version of the *Tripiṭaka Koreana* reproduced every glyph found in the blocks, making it more accurate for our purposes.
[3] English translation by Kieschnick (2014: 32).

ours, Liang et al. (2010) exploited syntactic parse trees and extracted the subject of the quotative verb as the speaker, but they reported no evaluation on their method's accuracy. We similarly make use of grammatical roles to identify speakers and listeners, using an algorithm with the following three main steps.

4.1 Extraction of utterances

We first extract utterances and the quotative verbs associated with them.

Quoted speech extraction. We extracted text enclosed within pairs of quotation marks, that is,「...」for Chinese. When a speech spans multiple paragraphs, an extra opening quotation mark is placed at the beginning of each paragraph. Not all quoted text, however, indicate direct speech; quotation marks may instead serve to highlight or emphasize words or phrases. We therefore included only quoted texts that end with a punctuation.

Quotative verb extraction. Direct speech is often associated with a quotative verb (e.g., "said", "told") whose subject and object indicate the speaker and listener. Typically, the quotative verb (e.g., 白 *bái* 'to address' in Figure 1) precedes the direct speech, which serves as its complement. As will be described in the next section, we identify the quotative verb by consulting the dependency parse tree of the sentence containing the direct speech.

For the portion of the Buddhist Canon with no quotation marks, we could not rely on quotation marks to identify the quotative verb. Instead, we manually compiled a list of the most frequent quotative verbs, and extracted all sentences that contain these verbs.

4.2 Speaker and listener attribution

In this step, we attribute a speaker and a listener to each utterance. Their identities are to be either extracted from the quotative verb, or inferred from context in a dialog chain.

From a quotative verb. Typically, the subject of the quotative verb, or its coordinated verb, is the speaker. For example, in Figure 1, "Ānanda" is the subject of the verb "prostrate", which are coordinated with the quotative verb *bái* 'to address'. The verb's indirect object is the listener. For instance, in Figure 1, "Buddha" is the indirect object of *bái*. There may be more than one speaker or listener for an utterance.

From a dialog chain. Because of radical pro-drop (Bisang, 2014), the subject of a sentence is often not mentioned in medieval Chinese (e.g., in Table 1). To mitigate this problem, we identify "dialog chains", where two characters take turns to speak. Such a chain usually has the format "X said ... Y replied ... X then said ...", and so forth. In this format, the quotative verb typically does not specify both the speaker and listener. Sometimes, there is no quotative verb at all.

Some dialogue chains are clearly marked, for example with the alternating sequence of 問 *wèn* ......答 *dá*...... *wèn* ...... *dá* .... In such chains, the speaker of an utterance is the listener of the previous or following utterance; and vice versa for the listener. Not all chains, however, follow this template. Whenever two utterances are sufficiently close, they can potentially belong to a dialog chain. The optimal threshold depends on the language and genre of the text. For example, Elson and McKeown (2010) used a distance threshold of 300 words for English. For the Buddhist Canon, after examining a set of dialog chains, we manually set the threshold to be 50 words. To validate a potential chain, we examine if the speakers and listeners, when explicitly mentioned, are swapped in the immediately preceding and following utterance. If so, we then infer from context the identities of the implicit speaker and listener of the utterance. Specifically, the listener is inferred to be the speaker of the previous or the following utterance. The speaker is assigned only if the same listener is specified in previous and following utterances.

4.3 Name standardization

A character may be mentioned in different ways, as a result of different combinations of his/her title and names (e.g., 文殊 "Mañjuśrī" and文殊菩薩 "Mañjuśrī Bodhisattva") and different epithets (e.g., the ten epithets of Buddha). We standardized character names using the *Buddhist Studies Person Authority Database* from Dharma Drum Buddhist College (DDBC, 2008), which contains entries for over 2000 characters in the Chinese Buddhist Canon and their alternative names.

5 Evaluation

To evaluate the quality of the network extraction method (Section 4), we constructed a gold-standard network for the book of *Ta ch'eng li ch'ü liu po lo mi to ching* 大乘理趣六波羅蜜多經 (henceforth to be referred to by its catalog number, K1381). An annotator manually identified the utterances, as well as their speakers and listeners. A total of 140 utterances were identified in this book, which contain 58,681 characters.

We first evaluated the precision and recall of our method in retrieval utterances. In other words, we compare the set of utterances retrieved by our method and those identified by the annotators, without considering the identities of the speakers and listeners. Our method achieved 96.0% precision at 85.0% recall for K1381.

We then measured the accuracy of our method in speaker and listener attribution. Among the correctly retrieved utterances in K1381, our method accurately identified 83.9% of the speakers and 84.7% of the listeners. Six of the errors were caused by two different dialog chains falling within the distance threshold, leading to incorrect inference of the implicit speakers and listeners. Other errors were caused by irregularities in the syntactic parse trees, for example when the quotative verb was misidentified. Finally, since the database of personal names (DDBC, 2008) is not exhaustive, some speakers and listeners were excluded from the network.

6 Network analysis

The method outlined in Section 4 extracted 506,848 utterances from the Buddhist Canon. We will focus our analysis on the top 100 characters with the most utterances. We first discuss our visualization of these utterances as a conversational network (Section 6.1), and then analyze the network with regard to the protagonist (Section 6.2) and the whole community (Section 6.3).

6.1 Network visualization

We visualize the utterances as a conversational network graph. In the graph, a node (or vertex) represents a person, i.e., a character in the text; and an edge from node X to node Y signifies that character X spoke to character Y[4]. Figures 2 shows the network graph for the Buddhist Canon[5]. The thickness of an edge is proportional to the number of utterances. The thicker the out-going edge from the node, the more the character spoke; and the thicker the in-coming edge, the more he or she listened.

[4] In conversational networks that use undirected edges (e.g., Elson et al., 2010), an edge indicates the presence of at least one instance of dialog interaction between the two characters, in either direction.

[5] The network graphs were drawn with the Graphviz library (Gansner and North, 2000).

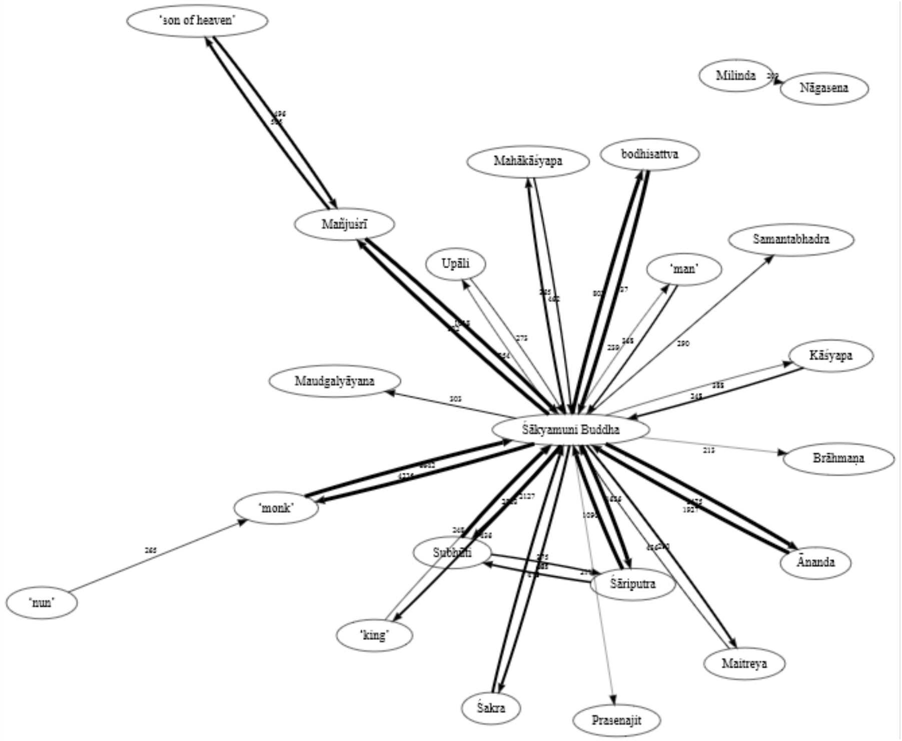

Figure 2. Conversational network of the Chinese Buddhist Canon, showing edges with 200 utterances or more.

6.2 Buddha and his interlocutors

Number of listeners. One metric for identifying the protagonists of a text is by out-degree centrality (Newman, 2010), which is defined as the number of out-going edges from a node; in other words, the number of other characters to whom the character spoke[6]. Buddha has a very high out-degree, talking to 95 of the 100 characters, far exceeding the median of 15 in the Buddhist network.

Number of utterances. Not only is Buddha well connected to the other characters, he also spoke more frequently. Figure 3 lists the top 10 speakers in the Buddhist Canon. Buddha's utterances constitute 44% of the total.

Frequent listeners. We now turn our attention to the characters to whom Buddha spoke. Buddha spoke most frequently to a group — the monks — rather than a specific individual (Table 2). He was especially close to some of his followers, and the utterance statistics bear out these special relationships. Buddha's closest companion is arguably his disciple and personal attendant, Ānanda, who accompanied him on his journeys for around thirty years, and listened to many of his sermons. After Ānanda, Buddha's most frequent listeners are two of his other disciples, Subhūti and Śāriputra. Subhūti is well known as the one to whom Buddha imparted many of his teachings, most famously in the *Diamond Sutra*. Śāriputra, called "General of the Dharma", is one of the two chief disciples of Buddha.

Beyond his disciples, Buddha's next most frequent conversational partner is Mañjuśrī, the embodiment of wisdom. Of the four great bodhisattvas, enlightened beings who have Buddha-nature, Mañjuśrī is regarded as the most important. The utterance statistics corroborate this assessment, placing him ahead of the other three great bodhisattvas, Maitreya, Samantabhadra and Avalokiteśvara.

[6] An alternative metric is the total utterance length of a character, which has the advantage of distinguishing short interactions from long ones, but can also be skewed by occasional long discourses.

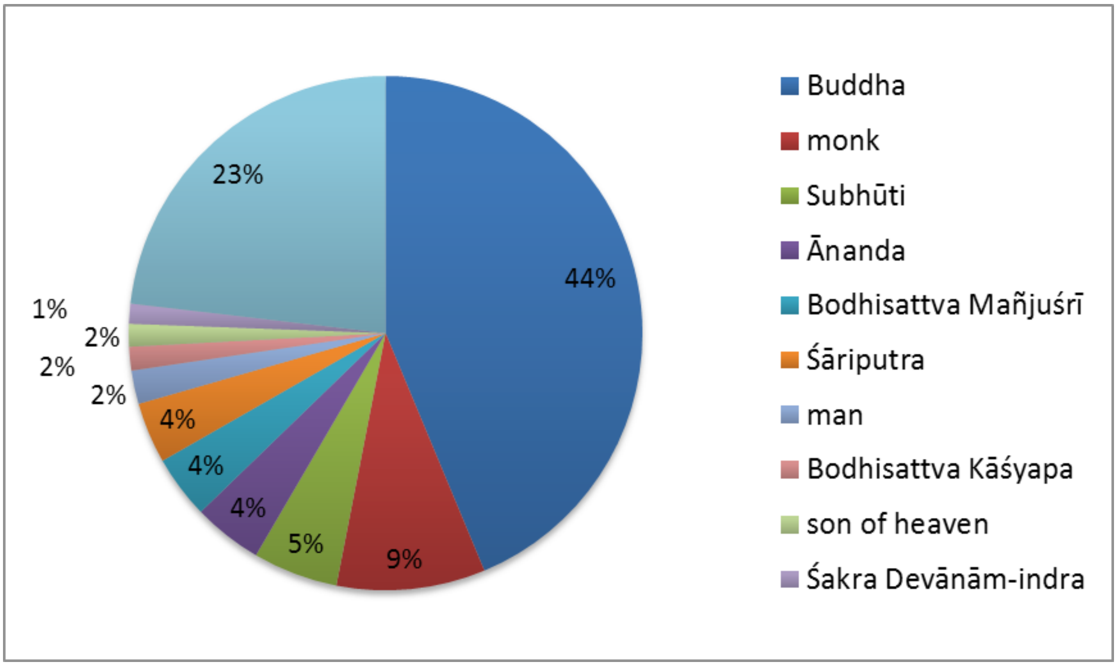

Figure 3. The ten most frequent speakers in the Chinese Buddhist Canon, in terms of number of utterances.

Table 2. The five characters to whom Buddha spoke most and listened most.

Spoke most to	Percentage	Listened most to	Percentage
monk	29.6%	monk	25.2%
Ānanda	15.3%	Subhūti	12.0%
Subhūti	9.4%	Ānanda	11.5%
Śāriputra	7.2%	Śāriputra	6.5%
Mañjuśrī	4.5%	Mañjuśrī	5.5%

6.3 Quotative verbs

In addition to his large number of listeners and utterances, Buddha also differs from other characters in terms of usage patterns in quotative verbs. Overall in the Buddhist Canon, the most frequent quotative verb is 言 *yán* 'to say' (20.6%), followed by 告 *gào* 'to tell' (13.4%) and 白 *bái* 'to address' (11.2%). However, as shown in Table 3, when Buddha spoke, he preferred *gào* (49.2%) over *yán* (30.1%) by a significant margin; and when Buddha listened, the speaker preferred *bái* (59.4%) overwhelmingly over *yán* (15.8%). What is more, Buddha never used *bái* when he spoke, and he was never addressed with *gào* when he listened.

Thus, the selection of *bái* and *gào* likely reflects not only his individual preference but an honorific usage; this usage can indeed also be detected in the rest of the network. The bodhisattvas, those closest to Buddhahood, spoke with *gào* and listened with *bái* to the disciples of Buddha[7], and even to the gods[8]. In turn, when the disciples of Buddha spoke to monks, who were relatively less close to the "Enlightened

[7] Among other examples, Śāriputra used *bái* when speaking to the bodhisattva Maitreya (e.g., 舍利弗白彌勒菩薩 [K0005]); and the bodhisattva Mañjuśrī used *gào* when speaking to Ānanda (e.g., 文殊師利告阿難言 [K0137]).

[8] Among other examples, both the Son of Heaven and the Brahma used *bái* when speaking to the bodhisattva Mañjuśrī and Sucintitārtha (e.g., 天子復白文殊師利 [K0224]) ; 爾時大悲思惟大梵天王 白海意菩薩言 [K1481])

One”, they similarly spoke with *gào* and listened with *bái* in the vast majority of the cases[9]. Thus, *bái* is reserved for speaking to someone of higher social status, and *gào* for speaking to someone of lower status. The usage patterns of quotative verbs can facilitate further research on the dynamics of the characters’ relationships — e.g., who tended to ask questions, and who tended to answer them; and on the hierarchy of the status/position of the characters (Lee and Wong, 2016).

Table 3. The most frequent quotative verbs among utterances where Buddha was speaker, or listener, respectively.

Buddha as speaker		**Buddha as listener**	
告[…言/曰] *gào [...yán/yuē]*	49.2%	白[…言/曰] *bái [...yán/yuē]*	59.4%
言 *yán*	30.1%	言 *yán*	15.8%
說 *shuō*	4.4%	問 *wèn*	4.3%
語 *yù*	4.0%	說 *shuō*	3.2%
問 *wèn*	2.7%	答言 *dáyán*	2.6%

6.4 Subcommunities

We now turn our attention to the network as a whole. As pointed out in Section 6.2, the network is centered on Buddha. Indeed, for 83% of the characters, their verbal interactions with Buddha represent over 90% of their total. For example, consider Subhūti and Śāriputra, two close friends especially known for their conversations in the Prajñāpāramitā Sutras. They have one of the strongest connections, with over 700 utterances between them. Still, Subhūti spoke and listened to Buddha (72.8% of his utterances) substantially more than to Śāriputra (13.7%); and Śāriputra likewise exhibited the same tendency (60.8% to Buddha and 17.4% to Subhūti). Indeed, among the character pairs with more than 150 utterances, 89% involve Buddha.

Nonetheless, there are few node clusters, or subcommunities, whose members connect with one another more than with Buddha. The most prominent ones include Nāgasena and King Milinda (K1002 *Miliṇḍapañha*), both of whom lived in the 2nd century BCE and therefore were not contemporaries of Buddha. Among the most frequent speakers, the Son of Heaven is the only one whose most frequent conversation partner was not Buddha; it was, rather, the bodhisattva Mañjuśrī. The high frequency of their conversation can be traced to a number of occasions when Buddha asked Mañjuśrī, one of the celestial bodhisattvas, to teach the Dharma to the Son of Heaven on his behalf.[10]

7 Conclusions

This article quantitatively analyzed verbal interactions among characters in the scriptures of Buddhism. We described a semi-automatic method to extract these interactions from the treebank of the Chinese Buddhist Canon, and evaluated its accuracy in attributing speakers and listeners to the utterances. We then visualized these interactions as a conversational network.

Our analysis has provided a quantitative view of Buddha as protagonist in the Canon, through the lens of his share of utterances and the number of characters to whom he spoke. Further, we discussed the types of characters with whom he interacted, and their use of quotative verbs, and the subcommunities.

Conversational networks complement more traditional methods for analyzing direct speech and present fresh views of a literary text. In future work, we plan to apply them to other works of literature and perform comparative analyses of their networks.

[9] Among other examples, Ānanda, Śāriputra and Subhūti all used *gào* when speaking to monks and nuns (e.g., 爾時尊者大目揵連告諸比丘 [K0648]; 時，舍利弗告諸比丘 [K0647]; 尊者難陁告諸比丘尼 [K0650])

[10] For example, in K44 and K222.

Acknowledgments: The work described in this article was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 155412).

References

Agarwal, Apoorv, Rambow, Owen, and Passonneau, Rebecca J. 2010. Annotation Scheme for Social Network Extraction from Text. In *Proc. Association for Computational Linguistics* (ACL).

Agarwal, Apoorv, Corvalan, Augusto, Jensen, Jacob, and Rambow, Owen. 2012. Social Network Analysis of Alice in Wonderland. *Proc. Workshop on Computational Linguistics for Literature*.

Baker, Paul, Gabrielatos, Costas, and McEnery, Tony. 2013. Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998-2009. *Applied Linguistics* 34(3):255-278.

Bingenheimer, Marcus, Hung, Jen-Jou, and Wiles, Simon. (2011). Social network visualization from TEI data. *Literary and Linguistic Computing* 26(3):271-278.

Bisang, Walter. 2014. On the strength of morphological paradigms: A historical account of radical pro-drop. In *Paradigm Change: In the Transeurasian Languages and Beyond*, pages 23–61.

Celikyilmaz Asli, Hakkani-Tur, Dilek, He, Hua, Kondrak, Greg, and Barbosa, Denilson. 2010. The Actor-Topic Model for Extracting Social Networks in Literary Narrative. In *Proc. NIPS Machine Learning for Social Computing Workshop*.

Chang, Pi-Chuan, Tseng, Huihsin, Jurafsky, Dan and Manning Christopher D. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proc. 3rd Workshop on Syntax and Structure in Statistical Translation*.

Crane, Gregory. 2006. What Do You Do with a Million Books? *D-Lib Magazine* 12(3). http://www.dlib.org/dlib/march06/crane/03crane.html

Csomay, Eniko. 2013. Lexical Bundles in Discourse Structure: A Corpus-Based Study of Classroom Discourse. *Applied Linguistics* 34(3):369-388.

DDBC. 2008. Buddhist Studies Person Authority Databases (Beta Version). Buddhist Studies Authority Database Project, Dharma Drum Buddhist College. Accessed at http://authority.ddbc.edu.tw/person/

Diesner, Jana, Frantz, Terrill L., and Carley, Kathleen M.. 2005. Communication Networks from the Enron Email Corpus: It's Always about the People, Enron is no

Different. *Computational and Mathematical Organization Theory* 11(3):201-228.

Doddington, George, Mitchell, Alexis, Przybocki, Mark, Ramshaw, Lance, Strassel, Strassel, and Weischedel, Ralph. 2004. The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation. In *Proc. Language Resources and Evaluation Conference* (LREC).

Elson, David K., Dames, Nicholas, and McKeown, Kathleen R. 2010. Extracting social networks from literary fiction. In *Proc. Association for Computational Linguistics* (ACL).

Elson, David K. and McKeown, Kathleen R. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (AAAI 2010), Atlanta, Georgia.

Gansner, Emden R., & North, Stephen C. 2000. An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, 30(11):1203-1233.

Holmes, David I. 1994. Authorship Attribution. *Computers and the Humanities* 28(2):87-106.

Hung, Jen-Jou, Bingenheimer, Marcus, and Wiles, Simon. 2010. Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations. *Literary and Linguistic Computing* 25(1):119-34.

Kieschnick, John. 2014. *A Primer in Chinese Buddhist Writings: Volume One: Foundations: Translation Key*. Department of Religious Studies, Stanford University. Accessed 18th August 2015. http://religiousstudies.stanford.edu/a-primer-in-chinese-buddhist-writings/

Knuth, Donald E. 1993. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Reading, MA: Addison-Wesley.

Lafferty, John, McCallum, Andrew, and Pereira, Fernando C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. International Conference on Machine Learning* (ICML), pages 282-289.

Lancaster, Lewis. 2010. From Text to Image to Analysis: Visualization of Chinese Buddhist Canon. In *Proc. Digital Humanities*.

Lancaster, Lewis and Park, Sung-bae. 1979. *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley: Berkeley University Press.

Lee, John and Kong, Yin Hei. 2016. *A dependency treebank of Chinese Buddhist texts.* In *Digital Scholarship in the Humanities* 31(1):140-151.

Lee, John and Wong, Tak Sum. 2016. *Hierarchy of characters in the Chinese Buddhist Canon*. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*, pages 531-534.

Liang, Jisheng, Dhillon, Navdeep, and Koperski, Krzysztof. 2010. A large-scale system for annotating and querying quotations in news feeds. In *Proceedings of the 3rd International Semantic Search Workshop*, pages 1–5.

Mahlberg, Michaela and Smith, Catherine. 2012. Dickens, the suspended quotation and the corpus. *Language and Literature* 21(1):51-65.

McDonald, Ryan, Lerman, Kevin and Pereira, Fernando. 2006. Multilingual dependency parsing with a two-stage discriminative parser. In *Proc. 10th Conference on Computational Natural Language Learning* (CoNLL-X).

Moretti, Franco. 1999. Atlas of the European Novel 1800-1900. London: *Verso*.

Moretti, Franco. 2011. Network Theory, Plot Analysis. *New Left Review* 68: 80-102.

Mutton, Paul. 2004. Inferring and Visualizing Social Networks on Internet Relay Chat. *Proc. 8th International Conference on Information Visualization*.

Newman, Mark. 2010. *Networks: An Introduction*. New York: Oxford University Press.

Oelke, Daniela, Kokkinakis, Dimitrios, and Keim, Daniel. A. 2013. Fingerprint Matrices: Uncovering the dynamics of social networks in prose literature. *Computer Graphics Forum* 32(3.4):371-380.

Pareti, Silvia, O'Keefe, Timothy, Konstas, Ioannis, Curran, James R., and Koprinska, Irena. 2013. Automatically Detecting and Attributing Indirect Quotations. In *Proc. Empirical Methods for Natural Language Processing* (EMNLP).

Pouliquen, Bruno, Steinberger, Ralf, and Best, Clive. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492.

Rydberg-Cox, Jeff. 2011. Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(3). https://letterpress.uchicago.edu/index.php/jdhcs/article/view/86

Sealey, Alison. 2010. Probabilities and Surprises: A Realist Approach to Identifying Linguistic and Social Patterns, with Reference to an Oral History Corpus. *Applied Linguistics* 31(2):215-235.

Stiller, James, Nettle, Daniel, and Dunbar, Robin I. M. 2003. The Small World of Shakespeare's Plays. *Human Nature* 14(4):397-408.

Xue, Naiwen, Xia, Fei, Chiou, Fu-dong, and Palmer, Marta. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11:207-238.

Zhao, Hai, Huang, Chang-Ning and Li, Mu. 2007. An Improved Chinese Word Segmentation System with Conditional Random Field. In H. T. Ng, & O. O. Y. Kwong (Eds.), *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, pages 162-165.