Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Hierarchy of Characters in the Chinese Buddhist Canon

John S. Y. Lee and Tak Sum Wong
Halliday Centre for Intelligent Applications of Language Studies
Department of Linguistics and Translation, City University of Hong Kong
tswong-c@my.cityu.edu.hk, jsylee@cityu.edu.hk

Abstract

With over 50 million characters in over 1500 texts, the Chinese Buddhist Canon is a complex literary collection. Besides the Buddha himself, there is a myriad of characters including bodhisattvas, deities, disciples of Buddha, monks, lay Buddhists as well as kings. This paper analyzes the hierarchy among these characters by examining their verbal interactions. Exploiting techniques from natural language processing, we extract all direct speech from the text, and examine the relation between the speakers, listeners, and the quotative verbs used for reporting the speech. We show that a number of the quotative verbs indicate relative status between the speaker and the listener. We then use their usage patterns to induce a hierarchy of the characters in the Canon.

Introduction

With over 50 million characters in over 1500 texts, the Chinese Buddhist Canon is a complex literary collection. Besides its protagonist, Buddha, there is a myriad of characters including bodhisattvas, deities, disciples of Buddha, monks, lay Buddhists, as well as kings. This paper analyzes the hierarchy among these characters by examining their verbal interactions. In particular, we investigate the relation between the speakers, listeners, and the quotative verbs (e.g., 'tell', 'say') used for report their dialogs. The sentence in Figure 1, for example, contains an utterance from Ānanda to Buddha, with 白 *bái* 'to address' as the quotative verb.

Exploiting techniques from natural language processing, we extract all direct speech from the Canon, and then analyze the distribution of the quotative verbs. We show that a number of these verbs, notably *gào* 'to tell' and *bái* 'to address', indicate relative status between the speaker and the listener. Finally, we use the usage statistics of these verbs to induce a hierarchy of the characters in the Canon.

Background

There is increasing interest in analyzing verbal interactions in literary texts. Dialogs between characters have been manually annotated for *Alice in Wonderland* (Agarwal et al., 2012) and parts of *The Story of the Stone* (Moretti, 2011), enabling studies on the protagonists and the characters associated with them. Mahlberg and Smith (2012) automatically extracted 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 applied it on 60 novels to investigate correlations between the number of characters, the amount of dialog interactions and the novel setting.

Numerous digital analyses on the Chinese Buddhist Canon have treated a wide range of research questions, from authorship (Hung et al., 2010), the origins of doctrinal terms (Lancaster, 2010), to relations between characters and locations (Bingenheimer et al., 2011). To the best of our knowledge, however, there is not yet any quantitative analysis on verbal interactions in the Canon.

Data

Textual material

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 (Lancaster and Park, 1979). 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 (CBETA).

Given the scale of our corpus, an automatic procedure is necessary to identify the speaker, listener, and quotative

verb for each utterance in the Canon. For example, from the utterance reported in the sentence in Figure 1, the algorithm must be able to identify Ānanda as speaker, Buddha as listener, and *bái* 'to address' as quotative verb.

Simple string search does not suffice since many characters have multiple meanings. For example, the character for *bái* also means 'white', and so a naive search would return many false positives. By performing part-of-speech (POS) tagging on the Canon, one can distinguish between the use of *bái* as verb and as adjective. However, POS tags alone are still inadequate, since they cannot indicate the speaker and listener. In Figure 1, while the listener (Buddha) immediately follows the quotative verb *bái*, some distance separates the verb from the speaker (Ānanda).

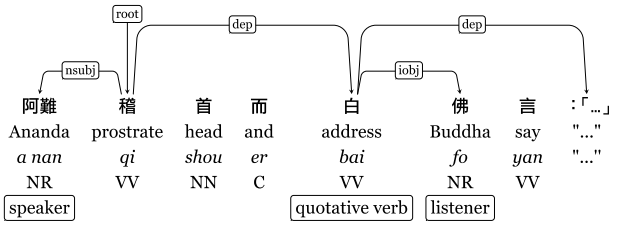

Figure 1. Parse tree for the sentence, "Ānanda prostrated and addressed Buddha, saying, '…'", showing part-of-speech tags for each Chinese word, and dependency relations that are used by our algorithm to identify the speaker, listener and quotative verb

Treebank

A treebank — a database of syntactic parses of each sentence in a corpus — provides the necessary syntactic information for our task. In a dependency treebank, every word is annotated with a part-of-speech tag and its dependency relation with its parent word. Figure 1 shows an example from a dependency treebank of Chinese Buddhist texts (Lee & Kong, 2014), which follows the POS tagset of the Penn Chinese Treebank (Xue et al., 2005) and the dependency labels from the Stanford Dependencies for Chinese (Chang et al., 2009). "Ānanda" is a proper noun (NR) which serves as the noun subject (nsubj) of the verb "prostrate" (VV). "Buddha", another proper noun, is the indirect object (iobj) of the verb "address".

Since this treebank covers only four sutras, we need to automatically derive parse trees for the rest of the Canon. Off-the-shelf Chinese syntactic parsers do not perform well on medieval Chinese, since they are trained on modern Chinese. Instead, using the treebank as training data, we built a word segmenter and part-of-speech tagger in the Conditional Random Fields (Lafferty et al., 2001) framework with the CRF^{++} implementation (Kudo, 2005). We then trained a Minimum-Spanning Tree parser (McDonald et al., 2006) to parse the rest of the Canon.

Data extraction

Given a parse tree, our algorithm first extracts the direct speech and its associated quotative verb, and then attributes a speaker and listener to the speech.

Direct speech and quotative verb extraction

Direct speech is enclosed within pairs of Chinese quotation marks, that is, 「… 」. It is often associated with a quotative verb (e.g., 'told') whose subject and object indicate the speaker and listener (e.g., *John* told *Mary*, "…"). In our corpus, the quotative verb (e.g., *bái* in Figure 1) usually precedes the direct speech, which serves as its complement. We extract all sentences with quotation marks, and then identify the quotative verb by consulting the dependency parse tree of the sentence.

Speaker and listener attribution

Typically, the speaker is the subject of the quotative verb or its coordinated verb, as is the case for "Ānanda" in Figure 1. The verb's object, indirect object ("Buddha") is the listener. We standardized character names using the *Buddhist Studies Person Authority Database* (DDBC, 2008), which contain entries for over 2000 characters in the Chinese Buddhist Canon and their alternative names.

Direct speech often takes the form of a "dialog chain", where character X and character Y take turns to speak. Such a chain usually has the format "X said … Y replied … X then said …", where the quotative verb typically does not specify both the speaker and listener. We considered two utterances that are sufficiently close[1] to belong to a dialog chain. Assuming that the speaker and listener of each utterance are swapped in the previous utterance, we inferred from context the identities of the implicit interlocutors.

To evaluate our data extraction algorithm, a human annotator identified the dialogs in *Ta ch'eng li ch'ü liu po lo mi to ching* 大乘理趣六波羅蜜多經 (K1381) Canon. The algorithm achieved 96.0% precision at 85.0% recall in retrieving the 140 utterances in this sutra. Among the correctly retrieved utterances, it was able to identify 83.9% of the speakers and 84.7% for listeners.

Analysis

We first discuss the distribution of quotative verbs in the Canon and for Buddha in particular. Next, we demonstrate that the choice of these verbs is correlated with one's status, and then use their usage statistics to induce a hierarchy for the characters.

[1] After examining a set of dialog chains, we empirically optimized the threshold to be 50 words for our corpus.

Quotative verbs

Table 1 shows the ten most frequent quotative verbs in the Buddhist Canon. The most frequent one is 言 *yán* 'to say'. This is a monotransitive verb, since it takes only one object, that is, the direct speech itself; it does not take the listener as an object. In contrast, the next two most frequent verbs, 告 *gào* 'to tell' and 白 *bái* 'to address', are both ditransitive. In addition to the direct speech, they also take the listener as an indirect object, such as "Buddha" in Figure 1. They are also optionally followed by 言 *yán* or 曰 *yuē*, which function like a participle.

Quotative verb	Frequency
言 *yán* 'to say'	20.6%
告 *gào* <listener> [言 *yán*/曰 *yuē*] 'to tell <listener> [saying]'	13.4%
白 *bái* <listener> [言 *yán*/曰 *yuē*] 'to address <listener> [saying]'	11.2%
說 *shuō* 'to say'	7.5%
答言/答曰 *dáyuē*/*dáyán* 'to reply and say'	7.1%
曰 *yuē* 'to say'	4.0%
問 *wèn* 'to inquire'	3.9%
語 *yǔ* 'to say'	2.3%
作 *zuò* 'to make'	1.5%
問曰 *wènyuē* 'to inquire and say'	1.5%

Table 1. Ten most frequent quotative verbs in the Chinese Buddhist Canon

Honorific use of quotative verbs

Buddha

Buddha's usage of quotative verbs diverges significantly from the overall statistics. Although the most frequent quotative verb is *yán* 'to say', when Buddha spoke, he preferred 告 *gào* 'to tell' over *yán* by a significant margin (49.2% to 30.1%; Table 2); and when Buddha listened, the speaker overwhelmingly preferred 白 *bái* 'to address' over *yán* (59.4% to 15.8%; Table 2). What is more, the "Enlightened One" never used *bái* in his more than 22000 utterances; and among the more than 16000 utterances to which he listened, he was never addressed with *gào*.

The non-collocation of *bái* with Buddha (as speaker) and *gào* (as listener) likely reflect not only individual writing style of the author or translator, but rather an honorific usage. It is well known that many Chinese words and phrases indicate social respect or deference. Studies on these honorifics tend to focus on expressions referring to oneself (e.g., 愚 *yú*) or to others (e.g., 陛下 *bìxià*); less attention has been paid to verbs. To test our hypothesis, we analyze their usage among other characters.

Buddha as speaker		Buddha as listener	
告 (…言/曰) *gào* 'to tell'	49.2%	白(…言/曰) *bái* 'to address'	59.4%
言 *yán* 'to say'	30.1%	言 *yán* 'to say'	15.8%
說 *shuō* 'to say'	4.4%	問 *wèn* 'to inquire'	4.3%
語 *yǔ* 'to say'	4.0%	說 *shuō* 'to say'	3.2%
問 *wèn* 'to inquire'	2.7%	答言/答曰 *dáyuē*/*dáyán* 'to reply and say'	2.6%

Table 2. The most frequent quotative verbs among utterances where Buddha was speaker, or listener, respectively.

Other characters

If a quotative verb indicates relative status between the speaker and listener, it should be used predominantly in one direction only between the two characters. Assuming character X and Y have different status, then the verb should be used *either* only when character X spoke to Y, *or* only when Y spoke to X. We thus measured how often each verb in Table 1 is used only in one direction between the two characters[2].

Two quotative verbs stood out. In 95.5% of the character pairs, the verb *bái* is used by one person to talk to the other, but not in the reverse direction. In 87.3% of the pairs, a similar trend held for *gào*. These figures suggest that the choice of *bái* and *gào* is strongly influenced by the identities of the speaker and listener[3]. More precisely, *bái* is reserved for speaking to someone of higher status, and *gào* for speaking to someone of lower status. One can thus induce a hierarchy of the characters in the Buddhist Canon by ordering them in a manner consistent with their *bái* and *gào* statistics.

Hierarchy among the characters

We ranked the top 100 characters in such a way as to minimize the number of "conflicts", i.e., the number of utterances where a higher-status character used *bái* when speaking to one with lower status, or where a lower-status character used *gào* when speaking to one with higher status. Out of more than 20,000 *bái* and *gào* utterances involving these characters, there are only 22 "conflicts"[4], suggesting that the hierarchy is well established.

Buddha naturally occupies the top spot. The bodhisattva Mañjuśrī (文殊) ranks second; he addressed everyone with *gào* except Buddha, and was addressed with *bái* by everyone, again except Buddha[5]. The rest of the hierarchy largely follows the major groups as listed below.

[2] We considered only character pairs who used the verb at least 5 times.
[3] The percentage for the other verbs are substantially lower, with *wènyuē* at 46.2%, and all others below 30%.
[4] Most of these conflicts result from inconsistent usage of *bái* and *gào* between two characters, e.g., the two bodhisattvas Vajrapani and Mañjuśrî, and the two disciples Mahākāśyapa and Śāriputra.
[5] On two occasions the bodhisattva Vajrapāṇi addressed Mañjuśrī with *gào*, e.g., 時金剛手菩薩復告文殊師利言 (K.1376).

Bodhisattvas

The bodhisattvas, or the "enlightened beings", were closer to Buddhahood than any of the groups below. This explains why they were respected by almost everyone else in terms of the quotative verb. All the 60 *gào* and *bái* utterances between a bodhisattva and a disciple conform to this expectation; for example, Mañjuśrī used *gào* when speaking to the disciple Ānanda[6].

Disciples

The disciples of Buddha consistently paid respect to the bodhisattvas; for example, Śāriputra used *bái* when speaking to the bodhisattva Maitreya[7]. However, when the disciples spoke to monks, who were less advanced on their way to Buddhahood, their verb usage pattern completely changed. There are plenty of examples where the disciples Ānanda, Śāriputra and Subhūti addressed monks and nuns with *gào*[8]. Among the *gào* and *bái* utterances between disciples and monks, more than 95% give the disciples higher status[9].

Deities

Deities reside in various realms in the Buddhist cosmology. Unlike the exalted and omnipotent gods in many other religions, they do not head the hierarchy in the Buddhist world: both the Son of Heaven (天子) and Brahmā (梵), for example, used *bái* when speaking to the bodhisattvas Mañjuśrī and Sucintitārtha[10].

The deities seem to be regarded as beneath not just the bodhisattvas but also the disciples; for instance, Śakra, the ruler of heaven, always paid respect to the disciples[11]. Few utterances, however, show how other deities related to them.

Kings

The two kings with most utterances are Ajātaśatru (阿闍世王) and Prasenajit (波斯匿王). Statistics with quotative verbs suggest that Ajātaśatru had lower status than the bodhisattvas and the disciples: for example, he always addressed Mañjuśrī with *bái*[12], and two disciples, Ānanda and Mahākāśyapa, addressed him with *gào*[13]. Prasenajit likewise addressed bodhisattva with *bái*[14], though the monks paid respect to him with *bái*[15]. It is difficult to generalize the trend to other kings, however, due to limited samples.

[6] E.g., 文殊師利告阿難言 (K.0137)
[7] E.g., 舍利弗白彌勒菩薩 (K.0005)
[8] E.g., 爾時尊者大目揵連告諸比丘 (K.0648); 時，舍利弗告諸比丘 (K.0647); 尊者難陁告諸比丘尼 (K.0650)
[9] Most exceptions involve the disciple Maudgalyāyana, e.g., 諸比丘告目連言 (K.0896).
[10] E.g., 天子復白文殊師利 (K.0224); 爾時大悲思惟大梵天王 白海意菩薩言 (K.1481)
[11] Śakra always used *bái* when speaking to Subhūti, such as 釋提桓因白須菩提 (K.0005); 天帝釋白善現言 (K.0001)
[12] E.g., 阿闍世王復白文殊師利 (K.0179)
[13] E.g., 爾時尊者阿難告阿闍世王言 (K.1483)
[14] E.g., 爾時波斯匿王白菩薩言 (K.0022)
[15] E.g., 時 比丘白王言 (K.0895)

Conclusion

We have examined the hierarchy of characters in the Chinese Buddhist Canon by analyzing the quotative verbs that report direct speech. We have shown the honorific usage of two of these verbs, *gào* 'to tell' and *bái* 'to address', and induced a hierarchy of the characters in the Canon on the basis of their usage patterns.

Acknowledgment. This work was partially supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. CityU 155412)

References

Agarwal, A., Corvalan, A., Jensen, J., and Rambow, O. 2012. Social Network Analysis of Alice in Wonderland. *Proc*. *Workshop on Computational Linguistics for Literature*.

Bingenheimer, M., Hung, J.-J., and Wiles, S. (2011). Social network visualization from TEI data. *Literary and Linguistic Computing*, 26(3):271-278.

Chang, P.-C., Tseng, H., Jurafsky, D. and Manning C. D. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proc*. *3rd Workshop on Syntax and Structure in Statistical Translation*.

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

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

Hung, J.-J., Bingenheimer, M., and Wiles, S. 2010. Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations. *Literary and Linguistic Computing* 25(1).

Kudo, T. 2005. CRF^{++}: Yet another CRF toolkit. Accessed at http://crfpp.sourceforge.net.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc*. *International Conference on Machine Learning* (ICML), pp.282-289.

Lancaster, L. 2010. Pattern Recognition and Analysis in the Chinese Buddhist Canon: A Study of "Original Enlightenment". Asia Pacific World 3rd series 60.

Lancaster, L. and Park, S. 1979. *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley: Berkeley University Press.

Lee, J. and Kong, Y. H. 2014. A dependency treebank of Chinese Buddhist texts. In *Literary and Linguistic Computing*.

Mahlberg, M. and Smith, C. 2012. Dickens, the suspended quotation and the corpus. *Langauge and Literature* 21(1):51-65.

McDonald, R., Lerman, K. and Pereira, F. 2006. Multilingual dependency parsing with a two-stage discriminative parser. In *Proc*. *CoNLL*.

Moretti, F. 2011. Network Theory, Plot Analysis. *New Left Review*, 68.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11: 207−238.