An Annotated Corpus of Direct Speech

John Lee, Chak Yan Yeung

Halliday Centre for Intelligent Applications of Language Studies
Department of Linguistics and Translation
City University of Hong Kong
E-mail: jsylee@cityu.edu.hk, chak.yeung@my.cityu.edu.hk

Abstract

We propose a scheme for annotating direct speech in literary texts, based on the Text Encoding Initiative (TEI) and the coreference annotation guidelines from the Message Understanding Conference (MUC). The scheme encodes the speakers and listeners of utterances in a text, as well as the quotative verbs that reports the utterances. We measure inter-annotator agreement on this annotation task. We then present statistics on a manually annotated corpus that consists of books from the New Testament. Finally, we visualize the corpus as a conversational network.

Keywords: direct speech, coreference, corpus annotation

1. Introduction

There is increasing interest in analyzing social networks in literary texts. Such networks have been constructed for *Alice in Wonderland* (Agarwal et al., 2012), *Les Misérables* (Newman and Girvan, 2004), *Biographies of Eminent Monks* (Bingenheimer et al., 2011), a set of British novels (Elson et al., 2010), *Hamlet* (Moretti, 2011) and various Classical Greek tragedies (Rydberg-Cox, 2011), to name just a few examples. Social networks can visualize the protagonists, closely related characters and communities in the text. They also support the analysis of text structures and properties, such as the perspective holder (Agarwal et al., 2012) and the density of dialog interactions (Elson et al., 2010).

While characters can socially interact in a variety of ways, their conversations — who talked to whom, and what they talked about — are key indicators of their relationships. This paper proposes a scheme for annotating direct speech in literary texts, based on the Text Encoding Initiative (TEI) and the coreference annotation guidelines from the Message Understanding Conference (MUC). The scheme encodes the speakers and listeners of utterances in a text, as well as the quotative verbs that reports the utterances. To the best of our knowledge, this is the first reported attempt of such an annotation task on a corpus of literary texts.

The rest of the paper is organized as follows. The next section presents the proposed annotation scheme. Section 3 reports inter-annotator agreement. Section 4 presents statistics on a manually annotated corpus that consists of books from the New Testament.

2. Annotation Scheme

We propose an XML annotation scheme that marks direct speech, the speakers and listeners, as well as the associated quotative verbs. The scheme aims to adopt the best practices from coreference annotation, and conforms to the standards of the Text Encoding Initiative (TEI) where possible. A summary of the elements and their attributes are provided in Table 1.

An utterance is a span of spoken words, typically within a pair of quotation marks. We enclose utterances with the TEI element `<said>`. The first `<said>` element in Figure 1, for example, marks the sentence "We have come from … with us" as an utterance. Our annotation excludes written communications, but includes ambiguous cases where the words might have been delivered verbally or in written form (e.g., "his wife sent a message to him: '…'"). Each `<said>` element may be associated with up to three pieces of information: the quotative verb, the speaker, and the listener.

2.1. Quotative Verb

An utterance is usually reported with a quotative verb. The verb typically appears in the phrase that precedes the utterance (e.g., "He *replied*, '…'"), or sometimes in the preceding sentence. It may appear in the phrase that follows the utterance (e.g., "'...,' he *replied*."); it may also interrupt the utterance (e.g., "'Very well,' he *replied*, 'but …'"). In other cases, there may be no such verb at all. We mark all quotative verbs with the `<quotative>` element and assign to each a unique ID. In Figure 1, there are two such verbs, both of which are "said".

For each utterance that is reported by a quotative verb, we indicate the verb's ID in the `quotative` attribute in the `<said>` element. For instance, the two utterances in Figure 1 are associated with their respective "said" through the IDs "VERB1" and "VERB2".

2.2. Speaker and Listener

We mark characters involved in speaking and listening with `<rs>`, the TEI element for referencing strings. Each `<rs>` element is given an ID. To indicate the speaker for a `<said>` element, we use the TEI attribute `who`. Its attribute value contains the ID of the `<rs>` element of the speaker. Imitating the attribute `who`, we propose the attribute `whom` to mark listeners. Both `who` and `whom` can take multiple IDs. For example, the first `<said>` element in Figure 1 has the attribute `who="#NAME2"`, which

points to the word "they" as its speaker (see next section for treatment of coreference); and the attribute whom="#NAME4 #NAME5", which points to "him" and "the men of Israel" as its listeners.

An utterance and its speaker and listener are not necessarily located close by. Consider the dialog chain: "Isaac said to his father Abraham, 'My father?' 'What is it, my son?' 'Here is the fire and the wood," Isaac said, 'but where is the lamb for the burnt offering?'" Although Abraham is the speaker and listener of the second and third utterance, he is not explicitly mentioned. The attributes who and whom attributes of these utterances, then, must reference the mention of "Abraham" which precedes the first utterance.

In other cases, the speaker or listener may be genuinely unknown. For example, a passive construction may not give information about the speaker (e.g. "*Joshua* was told, '…'") or the listener (e.g., "As spoken by the prophet *Isaiah*, '…'").

Elson et al. (2010) annotated an utterance only when the characters are mutually aware of each other and the speech is mutually intended for the other to hear. While most direct speech satisfy this criterion, we also accept a number of exceptions, such as monologues (e.g. "*God* said, 'Let there be light.'") and inanimate objects as listeners (e.g., "he said to *it*, '…'").

When <rs xml:id="NAME1" key="Gibeonites"> <COREF ID="1"> the residents of Gibeon </COREF> </rs> heard what Joshua did to Jericho and Ai, they did something clever. ...

<rs xml:id="NAME2"> <COREF ID="2" TYPE="IDENT" REF="1"> They </COREF> </rs> came to <rs xml:id="NAME3"> <COREF ID="3"> Joshua </COREF> </rs> at the camp in Gilgal and <quotative xml:id="VERB1"> said </quotative> to <rs xml:id="NAME4"> <COREF ID="4" TYPE="IDENT" REF="3"> him </COREF> </rs> and <rs xml:id="NAME5" key="Israelites"> the men of Israel </rs>, <said who="#NAME2" whom="#NAME4 #NAME5" quotative="#VERB1">"We have come from a distant land. Make a treaty with us."</said>

<rs xml:id="NAME6" key="Israelites"> The men of Israel </rs> <quotative xml:id="VERB2"> said </quotative> to <rs xml:id="NAME7"> <COREF ID="5" TYPE="IDENT" REF="1"> the Hivites </COREF> </rs>, <said who="#NAME6" whom="#NAME7" quotative="#VERB2">"Perhaps you live near us. So how can we make a treaty with you?"</said>

Figure 1: Example annotations of direct speech, taken from Joshua 9:3-7: "When the residents of Gibeon heard what Joshua did to Jericho and Ai, they did something clever … They came to Joshua at the camp in Gilgal and said to him and the men of Israel, 'We have come from a distant land. Make a treaty with us.' The men of Israel said to the Hivites, 'Perhaps you live near us. So how can we make a treaty with you?'" (NET, 2006).

Element	**Description**	**Attribute**	**Description**
<said>	Utterance	who	ID given to the <rs> element marking the speaker of the utterance
		whom	ID given to the <rs> element marking the listener of the utterance
		quotative	ID given to the <quotative> element marking the quotative verb that reports the utterance
<quotative>	Quotative verb	xml:id	Uniquely assigned ID
<rs>	Speaker or listener of an utterance	xml:id	Uniquely assigned ID
		key	Standardized name of the speaker or listener

Table 1: Elements and attributes in our proposed scheme. Definitions of the <COREF> element can be found in Hirschman and Chinchor (1997)

2.3. Coreference

The character mention is often neither a personal name nor a proper name. It may be realized as a pronoun (e.g., "they" in the first utterance in Figure 1) or as a referring expression (e.g., "the Hivites" in the second utterance, an alternative name for the Gibeonites). In our proposed annotation scheme, we mark not only the speakers and listeners but also their coreference chains. Specifically, we link character mentions to their antecedents using the annotation guidelines for MUC-7 (Hirschman and Chinchor, 1997). The compatibility allows, on the one hand, direct exploitation of corpora where coreference is already annotated under these guidelines; it also facilitates, on the other hand, use of our corpus as coreference data.

We annotate the character mention with a `<COREF>` element, with a `REF` attribute that points to a preceding `<COREF>` element with which it is coreferential[1]. For example, the speaker of the first utterance in Figure 1, "they", is marked by a `<COREF>` element with the attribute `REF="1"`. This attribute references the `<COREF>` element with the matching ID, i.e., "the residents of Gibeon". It is possible for the speaker and listener of an utterance to reference the same antecedent (e.g., "*They* began to discuss this among *themselves*, '…'"). We also allow an `REF` attribute to point to a `<COREF>` element positioned later in the text.

2.4. Standardized Names

The same character may have different mentions, as a result of name changes (e.g., "Abram" and "Abraham"), titles (e.g., "David" and "King David"), or alternative expressions (e.g., "residents of Gibeon" and "Gibeonites"). Conversely, different characters may have the same mention. For example, "John" can refer to one of Jesus' apostles, or the baptizer, or a relative of the high priest Annas, to count just a few possibilities.

To precisely identify the speakers and listeners, annotators may wish to establish a standardized set of names. In our annotation of the Bible (Section 5), for example, we decided to adopt the entries in two reference works, *Who's Who in the Old Testament* (Comay, 2001) and *Who's Who in the New Testament* (Brownrigg, 2001). Thus, the character "John" must be mapped to one of "John (son of Zebedee)", "John the Baptist", or "John (relative of Annas)". We specify the standardized name of each character marked by a `<rs>` element with the TEI `key` attribute. The character "residents of Gibeon" in Figure 1, for example, is mapped to "Gibeonites".

For many texts, it might be necessary to include common nouns in the list of standardized names to cover characters who are never referred to by personal names. For the Bible, for example, we included names for groups such as "scribes", the "chief priests", the "soldiers", etc. We also found it helpful to have a generic name, "individual", to capture all anonymous characters referred to as "a man", "someone", "a bystander", etc.

3. Inter-annotator Agreement

Two human judges — the instructor and teaching assistant of an introductory course for the Bible at a university — manually annotated the book of *Joshua* (18109 words) according to the scheme outlined above. They independently performed three tasks: identifying the utterances to be marked; attributing speakers and listeners to the utterances; and identifying the antecedent, if any, of the speakers and listeners. In this section, we report their level of agreement for these tasks. We did not measure agreement in mapping between the characters and the standardized names, since this task is generally unambiguous.

- ***Utterance identification***: One judge found 94 utterances in the book, while the other found 91. The disputed utterances are ambiguous as to whether they were spoken or written[2].
- ***Speaker and listener attribution***: Among the 91 utterances identified by both judges, they agreed on the speakers and listeners in all but one case: one judge selected "the family of Joseph" as listener, while the other selected "Ephraim and Manasseh", an appositional phrase.
- ***Coreference identification***: The two judges agreed perfectly on the antecedents of the speakers and listeners. Given the difficulty of the coreference task in general, this performance was somewhat surprising. In the context of the book of *Joshua*, this task was perhaps made easier by turn-taking of characters in conversations, and by repeated references to the same grounding instance.

4. Corpus Analysis

We applied the proposed annotation scheme on the four gospels — Matthew, Mark, Luke, and John — in the New Testament of the Bible.

Our manual annotation found 1,245 utterances, involving 148 characters. The vast majority of these utterances have both speakers and listeners; 8.4% have speakers only, and less than 0.5% have listeners only. The average utterance length is 44.8 words. The most frequent quotative verb is "said", accounting for 53.4% of the utterances; it is followed by "replied" and "answered".

In terms of the number of utterances, Jesus is the most frequent speaker and listener among the characters. Table 2 lists the five characters who listened most to Jesus. Not surprisingly, his disciples did so most frequently. Peter is ranked second, reflecting his leadership position among the disciples. The gospels

[1] E.g., Joshua spoke to "a man" (Joshua 5:13) who was revealed later to be "the commander of the Lord's army" (Joshua 5:15).

[2] E.g., "The king of Jericho received this report: '…'".

also frequently recorded Jesus' speech to the crowds, as well as his conversations with his main adversaries, the Pharisees and the scribes.

Listener	Percentage of utterances
Jesus' disciples	24.48%
Peter	7.29%
Pharisees	6.60%
crowds	6.05%
scribes	5.36%

Table 2: The top five listeners to Jesus in the gospels in terms of number of utterances.

The annotations in our corpus can be visualized as a conversational network (Figure 3). In this network 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. The thickness of an edge is proportional to the number of utterances. Hence, 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. All edges carrying eight or more utterances are shown.

This graph shows Jesus as the centre, as the protagonist in the text. His node has the highest out-degree; he spoke to 64.2% of the characters in the network. Ranked second is Peter, who spoke to only 10.4% of the characters. Excluding Jesus, the most frequent conversations occurred between Pontius Pilate and the crowds, reflecting the gospels' detailed portrayal of his trial of Jesus.

5. Conclusions and Future Work

We have presented a scheme for annotating direct speech in literary texts. The scheme combines best practices from the Text Encoding Initiative (TEI) and the coreference guidelines from the Message Understanding Conference (MUC). It encodes not only the utterances and their speakers and listeners, but also coreference chains and quotative verbs. We have shown that the annotation task can achieved high inter-annotator agreement.

Further, we have applied the annotation on the four gospels in the New Testament. We then presented a brief analysis of statistics of the direct speech therein, identifying the protagonist and his most frequent listeners. Finally, we visualized the annotated corpus as a conversational network.

In future work, we plan to expand our corpus to the rest of the Bible, with semi-automatic or automatic annotation (Elson et al., 2010), and to other literary works. We also intend to conduct further analyses on the annotated utterances to find distinctive vocabulary and speech styles of individual characters, as well as interaction patterns among the characters.

6. Bibliographical 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.

Brownrigg, R., & Brownrigg, C. R. (2001). *Who's who in the New Testament*. Psychology Press.

Comay, J. (2001). *Who's who in the Old Testament: Together with the Apocrypha*. Psychology Press.

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

Hirschman, L. and Chinchor, N. (1997). MUC-7 coreference task definition, version 3.0. *Proc. MUC-7*.

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

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2).

NET (2006). *The Net Bible*. Biblical Studies Press.

Rydberg-Cox, J. (2011). Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(3).

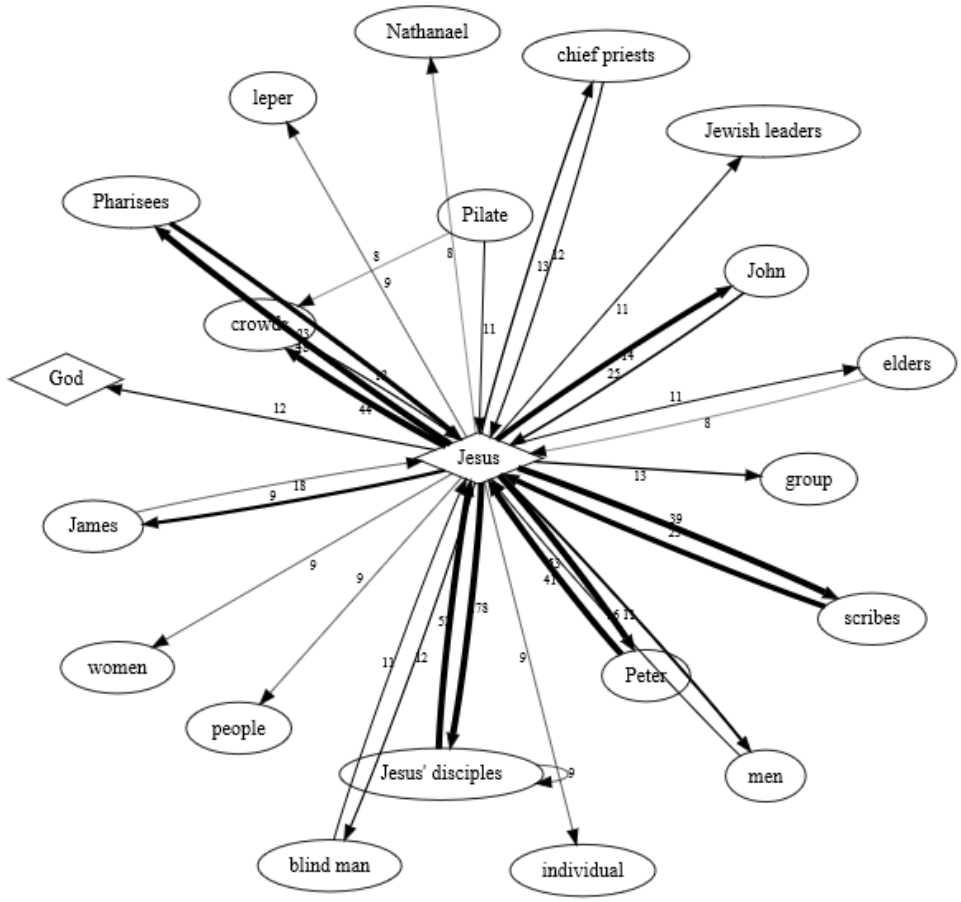

Figure 3: Conversational network generated from an annotated corpus of the gospels in the New Testament.