Annotation and Quantitative Analysis of Speaker Information in Novel Conversation Sentences in Japanese

Makoto Yamazaki, Yumi Miyazaki, Wakako Kashino
National Institute for Japanese Language and Linguistics
10-2, Midori-cho, Tachikawa City, Tokyo, Japan
{yamazaki, y-miyazaki, waka}@ninjal.ac.jp

Abstract

This study undertook a quantitative lexicological analysis using attributed speaker information, and reports on the problems of creating standards when annotating speaker information (gender and age) of conversation sentences in novels. In this paper, we performed a comparative analysis of vocabulary use by gender and age of conversation sentences and descriptive part sentences, as well as on the differences between Japanese novels and translations of foreign novels. In addition, a comparison with other spoken language materials was made.

Keywords: novels, conversation sentences, speaker information, gender, age, Balanced Corpus of Contemporary Written Japanese, characteristic words, descriptive part sentences.

1. Introduction

Although conversation sentences in novels are not actual speech, they were once used for language studies. This was probably because real spoken language data was difficult to obtain, unlike today. Language researchers of course recognized that conversation sentences in novels are not actual speech. For example, Takasaki (1981: 97) stated that "conversation sentences are difficult to use as speech materials and seem to require careful handling." Regarding conversation sentences in novels, Oishi (1987: 78) stated that "in summary, we conclude that conversation sentences are not useful as speech materials." Oishi (1987: 78-79) argued, however, that conversation sentences in novels can be considered representative of spoken language and be used for data in studies that understand the fundamental attributes of conversation.

We believe that conversation sentences in novels are effective for the investigation of the following research objectives:

(1) How are conversation sentences in novels different from actual conversation sentences? Clarifying these differences will make it possible to clarify the characteristics of actual conversations and can thus promote the understanding of the diversity of spoken language.

(2) A classical literary work such as *The Tale of Genji* and conversation sentences from a modern novel can be considered the same based on the structure of the novel. Therefore, diachronically studying conversation sentences and analyzing the changes make it possible to clarify historical changes in conversation sentences and the rhetoric used in them.

(3) By using speaker information, we can pursue the usage of role language (Kinsui, 2000), because role language appears in conversation sentences in novels more often than actual spoken language.

(4) Novels consist of conversation sentences and descriptive part sentences, and by analyzing the relationship between the two, we can expect that there are implications for literary studies and stylistic studies.

2. Data

The data used in this study were taken from novel samples[1] in the Balanced Corpus of Contemporary Written Japanese (BCCWJ). The BCCWJ includes 5,138 novel samples; however, the portion used in this study consisted of 2,419 samples taken mainly from the Library sub-corpus (LB), which completed the speaker information attribution work in July 2017. Out of this corpus, we targeted 2,088 samples in the novel genre that contained gender and age information.

The analysis language unit used in this study was the short-unit word[2] (SUW). Hereafter, one short-unit word will be expressed as "one word." In the analysis, in principle, parts of speech do not include punctuation marks, blank spaces, symbols, or unknown words.

3. Providing Speaker Information

Since speaker information is not provided in the BCCWJ, we created our own criteria and provided speaker information. The following tag information was used as clues:

<speech> tag: conversation sentences as block-level elements enclosed in parentheses.

<quote> tag: portions enclosed in parentheses on one line. In some cases, parentheses were used for non-conversation purposes, such as emphasis. Figure 1 shows a sample work file (Excel).

話者名	性別	年代	年代の確信レベル	非人間	会話モード	会話認定情報1	会話認定情報2	備考	職業	相手
井関弘志	男	成年層						２６歳	会社員	宮田良子／同僚
宮田良子	女	成年層						２５歳	組合の専従書記	井関弘志（同僚）

Figure 1. An example of providing speaker information (LBt9_00022).

[1] Whether a sample is a novel or not was determined by the NDC (book classification), which is meta-information of the sample. Specifically, the NDC regards texts that correspond to 9x3 (x=1 to 8) as novels.

[2] A short-unit word is a morphological unit annotated to the BCCWJ, which corresponds approximately to a dictionary entry. One SUW consists of, at most, two small, consecutive lexical items that carry meaning.

3.1 Annotations

Table 1 shows the annotations we provided, with 1–3 being required annotations. Annotations 4–9 were provided as necessary. Annotations 10 and 11 were not provided in any of the samples.

No	Attribute	Explanation
1	Speaker Name	Character's given name
2	Gender	Male, female, unknown
3	Age	Young (ages 0–19), adult (ages 20–59), elderly (ages 60 and over)
4	Confidence Level of the Age	Check if age estimation is extremely difficult.
5	Nonhumans	Check if there are beings besides humans, such as animals or spirits/ghosts.
6	Conversation Mode	Fill in for scenes that are not normal dialog scenes "telephone, dialect, foreigners, telepathy, quotes, soliloquys" and if conversation content is only symbols, representing doubt, silence, and surprise.
7	Conversation Certification Information 1	Fill in for speech that conforms to the format of a conversation, such as soliloquys, inner speech, etc.
8	Conversation Certification Information 2	Write down the basis for determining that this conforms to a conversation.
9	Remarks	General notes
10	Profession	Profession of the person making the utterance
11	Opponent	Conversation partner

Table 1. Speaker Attributes and Explanations

3.2 Criteria for Conversation Certification

In a novel, it may be difficult to judge whether certain text represents conversation. Therefore, Miyazaki (2017) created objective criteria and ensured that the annotation work did not become arbitrary.

Specifically, relevant parts were enclosed with “” and parts assumed to be someone actually speaking in the scene were regarded as conversation. Even if parts were not enclosed with “ ” , if it was determined that someone was actually speaking in the scene, it was considered conversation and that basis was recorded.

Thoughts, inner speech, conversations in dreams, and telepathy were also considered as forms of conversation and speaker information was given for these as well.

3.3 Problems during Annotation

・Conversation Certification

(1) In first-person narrated novels (i.e., novels that have a narrator who speaks and tells the story), we can also consider descriptive part sentences to be conversation sentences. The same is true for *rakugo* (Japanese sit-down comedy) transcription. The speaker information annotation policy for such cases has not yet been determined.

(2) There are assumed conversations, in which it is not clear whether someone spoke or not. For example, how would you handle a line such as, "A husband who comes back home and says nothing but 'eat, bathe, sleep.'"

・Speaker Information

(1) In novels, a speaker's gender is relatively easy to identify; however, it is rare that more than only approximate ages are known. Therefore, these cases cannot be categorized into detailed age groups. Thus, we must make a rough comparison when comparing with actual spoken language.

(2) When part of another person's quoted conversation is included in the conversation, the quoted part is treated without distinction. For example, if B's conversation is quoted in A's conversation, the speaker information is given to A only and B does not receive speaker information.

4. Results

4.1 Conversation Volume

In the 2,088 samples, the number of lines[3] certified as conversation totaled 207,071.[4] There were approximately 99.2 lines per sample. On the other hand, there were 274,66 lines of descriptive part sentence. [5] The ratio of conversation sentences to descriptive part sentences was approximately 0.754.

4.2 Distribution of Attributes of People Making the Utterances

Table 2 shows the distribution of utterances by gender and age. From Table 2, we can see that male utterances make up approximately 70% and adult (20‐59) utterances make up approximately 80%. Thus, novel conversations are primarily held by adult males.

	Young	Adult	Elderly	Unknown/ Other	Gender Total
Male	14,075 (6.8%)	119,507 (57.7%)	7,511 (3.6%)	1 (0.0%)	14,1094 (68.1%)
Female	13,232 (6.4%)	44,108 (21.3%)	3,070 (1.5%)	5 (0.0%)	60,415 (29.2%)
Unknown /Other	359 (0.2%)	4,964 (2.4%)	188 (0.1%)	51 (0.0%)	5,562 (2.7%)
Age Total	27,666 (13.4%)	168,579 (81.4%)	10,769 (5.2%)	57 (0.0%)	207,071 (100.0%)

( ) are percentage (%) of conversation parts overall

Table 2. Utterances by Gender and Age

[3] In the working file, the range of the <sentence> tag was one line, which we counted.

[4] Places where there was some information in both the gender and age fields.

[5] Places where there were blanks in both the gender and age fields.

Table 3 shows age and gender distribution for the calculated number of words based on short-unit words. We understand that if we compare the ratios, we find nearly the same numbers as in Table 2. This means that it is possible to estimate the approximate number of utterances (number of words) using the number of lines of utterances. With the usage of a simple calculation, there are approximately 9 words for each line of utterance.

	Young	Adult	Elderly	Unknown/Other	Gender Total
Male	99,394 (5.3%)	1,105,332 (59.3%)	72,224 (3.9%)	17,962 (1.0%)	1,294,912 (69.4%)
Female	87,287 (4.7%)	361,871 (19.4%)	26,965 (1.5%)	9,619 (0.5%)	485,742 (26.1%)
Unknown /Other	2,242 (0.1%)	50,013 (2.7%)	1,756 (0.1%)	30,252 (1.6%)	84,263 (4.5%)
Age Total	188,923 (10.1%)	1,517,216 (81.4%)	100,945 (5.4%)	57,833 (3.1%)	1,864,917 (100.0%)

Table 3. Conversation Volume by Gender and Age (number of short-unit words)

4.3 Differences between Conversation Sentences and Descriptive Part Sentences

Concerning the number of words, conversation sentences had a total of 1,864,917 words and 39,080 different words, and descriptive part sentences had a total of 4,211,887 words and 60,900 different words. In Guiraud's Index (R value), which is an indicator of vocabulary diversity, conversation sentences were 28.6, descriptive part sentences were 29.6. They showed almost the same value. We were surprised that conversation sentences were more redundant and had predicted that diversity would be much lower than descriptive par.

Table 4 shows the percentages of the parts of speech found in the conversation sentences and descriptive part sentences. Here we can identify some differences. Conversation sentences showed a lower percentage of nouns and a higher percentage of pronouns. Although numerically small, interjection values were 10 times higher in conversation sentences than in descriptive part sentences. Note that when we compared parts of speech with only the top 100 frequency words, there were 41 particle and auxiliary verbs in conversation sentences and 31 in descriptive part sentences. It can thus be said that there are more functional words in conversation sentences than in descriptive part sentences.

	Conversation Sentences		Descriptive Part Sentences	
	Number of words	Percentage	Number of words	Percentage
Nouns	416,799	22.35	1,184,428	28.12
Pronouns	70,406	3.78	75,641	1.80
Verbs	266,152	14.27	673,788	16.00
Adjectives	39,639	2.13	75,447	1.79
Adjectival nouns[6]	20,039	1.07	56,702	1.35
Adverbs	55,061	2.95	93,940	2.23
Adnominal adjectives	25,087	1.35	42,893	1.02
Conjunctions	6,263	0.34	13,244	0.31
Interjection	17,503	0.94	2,834	0.07
Particle	629,369	33.75	1,404,343	33.34
Auxiliary verbs	255,154	13.68	483,074	11.47
Prefix	16,846	0.90	16,548	0.39
Suffix	46,599	2.50	89,005	2.11
Total	1,864,917	100.00	4,211,887	100.00

Table 4. Percentages of Parts of Speech in Conversation Sentences and Descriptive Part Sentences

4.4 Differences between Japanese Novels and Translated Novels

Yamazaki (2017a) compared conversation sentences in Japanese novels with those in translated novels. His findings showed that, for example, in Japanese novels, words that are highly characteristic are suffixes and nouns attached to people, such as *san, sama, chan* (all are general suffixes of personal names but differ in politeness). On the other hand, in translated novels, many pronouns and proper nouns are highly characteristic. The greater use of personal pronouns in translated novels than in Japanese novels is thought to be due to the influence of translation from languages that use many personal pronouns, such as English. Among elderly people in Japanese novels, it was pointed out that *role language*[7] such as *washi* (i.e., a first-person pronoun used for elderly people) and *no* (i.e., a final particle used for elderly people) were seen. These examples of role language were also found in translated novels[8] but were used less than in Japanese novels. We hypothesized that since there is no equivalent to role language for elderly people in English, the translators might make greater use of these stereotypical words for describing people in a lively way. We still do not have a clear explanation for this, but we assume that the characters in many Japanese historical novels (i.e., *samurai* novels) would increase the use of such role language.

Table 5 shows how word type usage is distributed by gender. All the words–except those used by characters with unknown genders–were classified into three categories, which are words used only by females (female only), words used only by males (male only), and words used by both (common). From Table 5, we can see that half of the words are used by male characters only in both novels, while the number of words used by female characters only are relatively few. The ratio of common words in translated

[6] "Adjectival nouns" stem from what is known as adjective verbs in school grammar. For instance, *Tokubetsu* (special), *tashika* (sure, certain), *juyo* (important).

[7] Role language is the stereotypical use of words that reminds us of particular people, such as women, children, old people, etc.

[8] For example, *washi* was used in the translations of novels by Charles Dickens, Mark Twain, Maurice Leblanc, Victor Hugo, Jules Verne, Dickson Carr, Agatha Christie, etc.

novels is almost 10 points higher than that of Japanese novels.

	Japanese Novels	Translated Novels
Female only	3,455 (14.1%)	1,670 (9.6%)
Male only	18,054 (54.2%)	6,406 (50.2%)
Common	14,455 (31.6%)	3,736 (40.2%)
Total	11,812 (100%)	35,964 (100%)

Table 5. Distribution of Word Types by Gender

4.5 Comparison with Other Registers

In this section, we compare the characteristics of novel conversation sentences with other spoken language materials. Yamazaki's (2017b) comparison of the Corpus of Spontaneous Japanese (CSJ)'s Academic Presentation Speech (APS) and Simulated Public Speaking (SPS), Corpus of Everyday Japanese Conversation (CEJC), Nagoya University Conversation Corpus (NUC), Women's Words/Men's Words (Workplace Edition) (Workplace), and the BCCWJ's novel conversation sentences (Novels) lead to several conclusions.

Figure 2 shows the ratios of the three distinctive and frequently used parts of speech in spoken words, which are final particles, interjections, and filled pauses. From the ratios of those parts of speech, we can group the corpora into three categories: (1) SPS and APS, (2) NUC, Workplace, and CEJS, and (3) novels. This categorization corresponds to dialog, conversation, and written conversation, respectively.

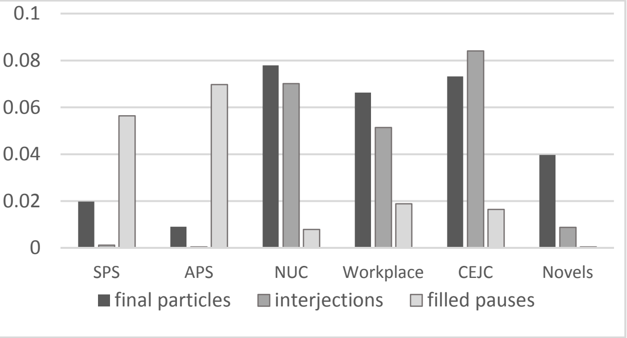

Figure 2. Ratio of final particles, interjections, and filled pauses.

The analysis results of characteristic words based on the Log-Likelihood Ratio are as follows. In simulated public speaking, filled pauses such as *ano*, *maa*, and *nn*, occupied the top positions. However, in academic public speech, even with the same filled pauses, filled pauses expressing mental operations taking time such as *eh* and *eh-to* (Sadanobu & Takubo, 1995) were characteristic words. In the NUC, final particles such as the interjections *uhn*, *neh*, and *sa* appeared in the top position. In workplace conversation, filled pauses did not appear; however, interjections such as *hai*, *ah*, and *eh* occupied the top positions. The greeting "good morning" was also characteristic of workplace conversations. In everyday conversations, words used for responses such as *uhn*, *u-uhn*, and *so* were in the top positions. The novel conversation sentences had completely different aspects, and pronouns such as "you," "I" (male), "I" and "you" (male) were characteristic words. It is also noteworthy that final particles, which represent role language, such as *wa*, *yo*, and *zo* were in the top positions.

5. Future Tasks

We are planning to release the speakers' information (gender and age) in our web concordancer "*Chunagon,*" which is free to applicants in 2019. *Chunagon* is the most frequently used web concordancer in the Japanese corpus. It has over 5,000 users. Other speakers' information will be released separately after several confirmations.

6. Acknowledgements

This work was supported by the NINJAL institute-based project "A Multifaceted Study of Spoken Language Using a Large-Scale Corpus of Everyday Japanese Conversation" (project leader: Hanae Koiso) and by JSPS KAKENHI Grant Number JP12345678.

7. Bibliographical References

Kinsui, S. (2000) Proposal of role language. In Sato, K. (ed) *Kokugo Ronkyu* 8 *New Aspect of Historical Japanese*, Tokyo: Meiji Shoin.

Miyazaki, Y., Kashino, W., and Yamazaki, M. (2017) Fundamental planning of annotation of speaker's information to utterances: Focused on novels in "Balanced Corpus of Contemporary Japanese," *Language Resource Workshop 2016*, Tokyo: National Institute for Japanese Language and Linguistics.

Oishi, H. (1987) Value of conversation sentences in modern novels as data of linguistic research, Kokubungaku Kaishaku to Kansyo, 52(7), pp.72-79.

Sadanobu, T. and Takubo, Y. (1995) The monitoring devices of mental operations in discourse-- a case of "eeto" and "ano(o)"--, *Gengo Kenkyu*, 108, pp.74-93.

Takasaki, M. (1981) On the conversation sentence in novels, Kotoba 2, pp.86-97.

Yamazaki, M. (2017a) Vocabulary diversity in conversation as seen in different corpus registers, *Language Resource Workshop 2017*, Tokyo: National Institute for Japanese Language and Linguistics.

Yamazaki, M. (2017b) Quantitative comparison of Japanese novels and translated novels with respect to lexicology, *Goi Kenkyukai 2017*, Sep. 16 at Aichi Gakuin University Sakae Satellite.

8. Language Resource References

National Institute for Japanese Language and Linguistics (2011) Balanced Corpus of Contemporary Written Japanese (BCCWJ).
http://pj.ninjal.ac.jp/corpus_center/bccwj/en/