Hindawi
Advances in Human-Computer Interaction
Volume 2018, Article ID 5056268, 9 pages
https://doi.org/10.1155/2018/5056268

Research Article

Factor Analysis of Utterances in Japanese Fiction-Writing Based on BCCWJ Speaker Information Corpus

Hajime Murai

Department of Complex and Intelligent Systems, Future University Hakodate, Hakodate 041-8655, Japan

Correspondence should be addressed to Hajime Murai; h_murai@fun.ac.jp

Received 30 May 2018; Accepted 28 October 2018; Published 19 November 2018

Guest Editor: Akinori Abe

To analyse the characteristics of utterances in Japanese novels, several attributes (e.g., the speaker, listener, relationship between the speaker and listener, and gender of the speaker) were added to a randomly extracted Japanese novel corpus. A total of 887 data sets, with 5632 annotated utterances, were prepared. Based on the attribute annotated utterance corpus, the characteristics of utterance styles were extracted quantitatively. A chi-square test was used for particles and auxiliary verbs to extract utterance characteristics which reflected the genders of and relationships between the speakers and listeners. Results revealed that the use of imperative words was higher among male characters than their female counterparts, who used more particle verbs, and that auxiliaries of politeness were used more frequently for 'coworkers' and 'superior authorities'. In addition, utterances varied between close and intimate relationships between the speaker and listener. Moreover, repeated factor analyses for 7576 data sets in BCCWJ speaker information corpus revealed ten typical utterance styles (neutral, frank, dialect, polite, feminine, crude, aged, interrogative, approval, and dandy). The factor scores indicated relationships between various utterance styles and fundamental attributes of speakers. Thus, results of this study would be utilisable in speaker identification tasks, automatic speech generation tasks, and scientific interpretation of stories and characters.

1. Introduction

To process story texts automatically using information technologies and artificial intelligence, it is necessary to identify the relationships between linguistic characteristics and attributes in the story. Writing styles are affected by various attributes such as genre, time and culture settings, social backgrounds, personalities of the characters, and the mood of a scene. Those characteristics of written styles have been utilised for text categorization and author identification tasks [1, 2]. However, various complex components within those styles have not been investigated enough individually except for a few aspects of gender or age [3–5]. If relationships between the words and profound concepts in story texts are identified, algorithms which interpret stories as flexibly as human beings may be developed. Moreover, such mechanisms would be conversely applicable to automatic story generation systems.

Among the various elements in story texts, conversational sentences are a challenge for automatic processing. Colloquial language often includes irregularities, reflecting daily usage of omissions and idiomatic expressions. Therefore, it is difficult to process irregular word sequences using natural language processing techniques.

Moreover, in novels or general story texts, each character is differentiated based on their manner of speech; it is a popularly used technique to help readers understand each character's personality [6]. Some readers can identify various attributes (e.g., gender, age, temperament, and social status as in real daily conversations [7, 8]) of characters in a text based on the characteristics of each character's dialogues. Moreover, even if the speaker's identity is not elaborated through descriptive sentences, most readers can accurately identify the speaker through conversational sentences. The distinct stylistic characteristics in each speaker's manner of speech tend to be exaggerated in conversations between fictional characters (particularly in the entertainment content). Therefore, although these characteristics do not precisely reflect the conversational styles of real people [9], they seem to function effectively as common symbols between the writers and readers of fictional texts. Thus, these characteristics form one of the cultural styles in story texts.

Previous research on the characteristics of distinct conversational styles has not only clarified the types of implied attributes of story characters but also investigated the historical and cultural origins of those styles [6]. However, these results are based on individual interpretations of the researchers. In addition, researchers have performed evaluation experiments to identify the characteristics of speakers based on sample sentences of spoken languages [10, 11]. The purpose of these experiments has been to facilitate automatic speech generation for interactive dialogues. The characteristics of speech style, depending on the age and personality of the speaker, have been analysed through psychological experiments.

However, no empirical evidence has established which type of distinct conversational style implies which type of attribute among fictional characters based on large-scale corpus.

If the characteristics of the distinct conversational styles of fictional characters could be quantitatively extracted from story texts using the methods of digital humanities, it would be possible to scientifically analyse the narratological functions and personal attributes of fictional characters. Results of the scientific analyses would provide objectification and falsifiability to the interpretation of narratives. Moreover, it would clarify effective features for identifying characters' personality. Therefore, it would become useful information in order to solve speaker identification problems in natural language processing based on the relationships between attributes of fictional characters and conversational styles.

2. Materials and Methods

To analyse relationships between attributes and conversational sentences, a tagged dialogue corpus of Japanese novels was employed [12]. This corpus is based on a random sampling of Japanese novel texts within the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [13]. The Japanese novel texts, included in the Nippon Decimal Classification class number 913, were extracted from the library-based corpus in the BCCWJ, and 100 texts were randomly selected (the appendix). Although there is also BCCWJ speaker information corpus which covers all Japanese novel texts in BCCWJ, that corpus includes only gender and age attributes.

Conversational sentences within the selected texts were extracted and attributes of the speaker (name, gender, occupation) and listener (name), relationship between the speaker and listener, and situations (e.g., family, office, criminal investigation) were manually added to each utterance. A total of 5632 utterances from 100 Japanese novel texts were tagged. Utterances with common attributes (i.e., same speaker and listener) were integrated as one data set and 887 data sets were obtained.

Tables 1 and 2 categorically describe the number of attributes for 887 data sets. Table 1 indicates that most of the speakers in Japanese novels are male characters. Table 2 shows the frequently appearing relationship attributes between the speaker and listener. The relationships between the speaker and listener are not always clearly described in story texts. Therefore, in the developed corpus, 484 of the 887 data sets did not have relationship attributes.

Table 1: Number of gender attributes for the data sets.

Male	510
Female	232
Other	145

Table 2: Number of relationship attributes for the data sets.

Friend	84
Co-worker	44
Subordinates	30
Superior authorities	29
Enemy	27
Brother, Sister	20
Spouse	19
Lover	17

Moreover, in order to perform statistically a factor analysis for frequencies of particles and auxiliary verbs, all Japanese literature texts in BCCWJ speaker information corpus have been utilised. BCCWJ speaker information corpus includes 11860 data sets of each character's utterances. However, due to statistical limitations, 7576 utterance data sets with total frequencies higher than 20 were selected. Because of limitation of included attributes, analysis about factor scores have been done only for gender and age attributes.

3. Results and Discussion

3.1. Characteristics of Text Styles. In this study, frequencies of functional words in utterances were selected as characteristics of text style since, in many Japanese novels, different usage patterns of functional words are used to indicate characters' personality [6].

In Japanese language, functional words mainly correspond to particles and auxiliary verbs. Therefore, statistical significance of the frequencies of particles and auxiliary verbs was analysed using a chi-square test.

The BCCWJ provides morphologically analysed data sets for the included novel texts. Therefore, particles and auxiliary verbs in utterances were extracted and counted by 887 data set units.

Table 3 presents the results of the chi-square test to identify the frequently appearing particles and auxiliary verbs for two categories of gender attributes (male and female); the statistically significant frequently used particles and auxiliary verbs ($p = < 0.001$) for each gender are presented in the table.

The utterance styles of 'male' included more imperative words (na, zo, and ya), whereas those of 'female' included particle words (no, yo, wa, kashira, and mono), implying a soft and feminine tone. These results are consistent with the general characteristics of feminine or masculine utterances [14].

Table 3: Significant words for each gender.

	Significantly more	Significantly less
Male	*ha, no* (case particle), *wo, to, ka, noda, na, zu, ga, toiu, zo, ya*	*te, no, nai, yo, ne, nodesu, teru, wa, kashira, mono*
Female	*te, no, nai, yo, ne, nodesu, teru, wa, kashira, mono*	*ha, no* (case particle), *wo, to, ka, noda, na, zu, ga, toiu, zo, ya*

Table 4: Significant words in division of relationships between speakers and listeners.

	Significantly more	Significantly less
Friend	*da, no, yo, ne*	*ha, mo, masu, desu, zu*
Co-worker	*masu*	*mo*
Subordinates	*ha, wo, to, ka*	*te, masu, desu, yo, ne, kara*
Superior authorities	*masu, desu, zu*	*da, yo, nai, na*
Enemy	*wo, na*	*masu*
Brother, Sister	*te*	*da, ne*
Spouse	*zu*	
Lover	*te, yo, ne*	*zu, na*

Similar to Table 3, Table 4 describes the eight frequently noted categories of relationships between the speaker and listener. Auxiliary verbs of politeness (masu and desu) were found to be used for 'coworkers' and 'superior authorities'; however, they were less frequently used for 'friend', 'subordinates', and 'enemy'. Moreover, although the reasons may be different, there is no need to express feelings of respect, which are meant for superiors, towards 'friend', 'subordinates', and 'enemy'. Relationship between 'friends' is that of equals in most cases. Fictional superiors generally do not express feelings of respect towards their subordinates. In addition, it is not usual for a person to politely speak to one's enemies.

Both 'friend' and 'brother, sister' were identified as close relationships; however, the characteristics of utterance styles were completely different. The use of da and ne was significantly more for 'friend' and significantly less for 'brother, sister'. 'Friend' typically implies an intimate person, while some people may dislike their 'brother, sister'; stories often depict complex relationships among siblings (e.g., Cain and Abel). If someone dislikes one's friend, the friendship ends. However, if someone dislikes one's brother, their relationship still remains. Therefore, these opposite characteristics may signify the difference between close and intimate relationships.

Moreover, an interesting observation was made regarding the characteristics of utterances for 'spouse' and 'lover.' The frequently used word for 'spouse' was the negative zu; however, for 'lover', words with intimate tones (te, yo, and ne) were more significant and negative words (zu and na) were less significant. Novels often depict matrimonial conflicts; therefore, this result may also reflect the characteristics of stereotypical fictional characters.

3.2. Factor Analysis for Utterance Styles. To extract the typical utterance styles of Japanese novel characters, a factor analysis for frequencies of particles and auxiliary verbs was performed. In order to extract statistically comprehensive utterance style, utterance sentences from all Japanese literature texts in BCCWJ speaker information corpus were utilised. Those utterance data possess three fundamental attributes (speaker names, genders, and ages) for each utterance sentence, although detailed attributes were not given. Each data for factor analysis is a 100-dimensional vector for one fictional character's all utterances. Those dimensions indicate frequencies of 100 types of frequently appearing particles and auxiliary verbs. Due to statistical limitations, 7576 utterance data sets with total frequencies higher than 20 were selected from 11860 data sets. The Promax rotation method was used and a parallel analysis was performed to determine the number of factors. After the factor analysis, less significant words (with a maximum factor loading of > 0.4) were eliminated and a subsequent factor analysis was performed repeatedly. Finally, after performing the factor analysis four times, ten factors were identified. The resultant factor scores are shown in Table 5; the bold font signifies cells whose factor scores exceeded 0.4.

Ten factors corresponded with the frequently appearing utterance patterns in Japanese novels. The characteristics and naming of each factor are as follows:

(i) Factor 1: It included the most frequently used neutral particles and auxiliary verbs. However, Factor 1 did not include words which indicated specific attributes; in other words, it represented a 'neutral style' of utterance.

(ii) Factor 2: This factor included friendly and frank particles and auxiliary verbs (e.g., tte, chau, teru, nanka, mono, yo, and mitai). Therefore, Factor 2 was referred to as 'frank style'

(iii) Factor 3: This factor included many words which were characteristically used in various Japanese dialects (e.g., ya, hen, da, nen, and haru). Therefore, Factor 3 was referred to as 'dialect style'.

(iv) Factor 4: This factor included formal and polite auxiliary verbs (e.g., masu, desu) and was referred to as 'polite style'.

Table 5: Factor loadings of influential, frequently appearing particles and auxiliary verbs.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
No (Case particle)	**1.03**	0.01	0.02	-0.02	-0.02	-0.03	0.02	-0.14	0.03	0.02
Wo (Case particle)	**1.03**	-0.11	-0.01	-0.05	0.04	-0.02	-0.01	-0.02	-0.02	0.01
Ni (Case particle)	**0.99**	0.04	0.00	0.01	-0.02	-0.02	-0.01	-0.01	-0.06	0.00
Kara (Case particle)	**0.94**	0.10	0.02	-0.05	-0.03	-0.03	-0.02	-0.18	0.01	-0.04
Ta (Auxiliary verb)	**0.90**	0.06	0.02	-0.02	0.03	-0.02	-0.04	0.01	0.05	-0.05
Ha (Binding particle)	**0.88**	-0.12	0.00	0.00	0.01	-0.05	0.01	0.14	0.05	0.05
Ga (Case particle)	**0.88**	0.08	0.00	0.00	-0.01	-0.02	0.00	0.04	0.06	0.00
Te (Conjunctive particle)	**0.86**	0.02	0.00	0.03	0.02	0.02	0.01	0.11	0.05	-0.07
Reru (Auxiliary verb)	**0.85**	0.03	-0.02	0.09	-0.05	-0.05	-0.01	-0.17	-0.02	0.04
To (Case particle)	**0.81**	-0.02	-0.01	0.06	-0.04	-0.02	0.03	0.10	0.13	-0.04
Da (Binding particle)	**0.72**	0.08	-0.04	-0.18	-0.05	0.00	-0.04	0.21	0.05	0.25
De (Case particle)	**0.70**	0.17	0.03	0.06	-0.02	0.01	0.04	0.11	0.10	-0.03
Rareru (Auxiliary verb)	**0.66**	0.01	-0.03	0.02	-0.01	-0.01	-0.02	0.03	-0.15	-0.04
Seru (Auxiliary verb)	**0.65**	-0.06	0.01	-0.07	0.06	0.04	-0.01	-0.07	-0.05	0.05
Mo (Binding particle)	**0.65**	0.08	0.02	0.06	0.03	0.02	0.06	0.31	0.00	-0.06
Made (Adverbial particle)	**0.64**	0.05	0.00	0.01	-0.02	0.02	-0.01	0.02	-0.07	0.01
Ba (Conjunctive particle)	**0.60**	-0.07	-0.06	0.06	0.01	-0.10	0.08	0.17	-0.15	0.11
Nai (Auxiliary verb)	**0.60**	0.08	-0.08	-0.15	0.08	0.03	-0.10	**0.40**	-0.05	0.00
Ga (Case particle)	**0.57**	-0.28	0.00	0.15	-0.06	-0.04	0.03	-0.03	0.24	0.23
To (Conjunctive particle)	**0.53**	0.05	-0.01	0.04	-0.03	0.01	0.04	0.09	0.22	-0.08
He (Case particle)	**0.49**	-0.08	0.03	0.01	0.03	0.07	0.04	-0.07	0.07	0.00
Shika (Adverbial particle)	**0.46**	0.07	0.00	0.04	-0.03	-0.02	-0.05	0.03	-0.08	-0.01
Dake (Adverbial particle)	**0.45**	0.06	0.07	-0.02	0.00	-0.02	-0.01	0.23	0.01	0.07
Ka (Adverbial particle)	**0.45**	0.09	0.00	0.02	0.00	0.03	0.00	0.38	0.08	-0.04
Nagara (Conjunctive particle)	**0.44**	-0.04	-0.01	0.01	0.06	0.07	0.02	0.04	-0.04	-0.04
Tte (Adverbial particle)	-0.01	**0.87**	0.02	0.00	-0.08	0.09	-0.01	0.01	-0.05	0.07
Chau (Auxiliary verb)	-0.08	**0.78**	-0.02	0.06	-0.04	-0.01	0.11	-0.12	0.01	0.02
Teru (Auxiliary verb)	0.06	**0.67**	0.08	-0.06	-0.03	0.04	-0.08	-0.03	0.01	0.21
Keredo (Conjunctive particle)	0.09	**0.61**	0.07	-0.01	-0.02	-0.04	0.00	0.14	0.13	-0.12
Nanka (Adverbial particle)	0.00	**0.52**	-0.01	0.03	-0.08	0.01	0.02	0.07	0.10	0.03
Mono (Ending particle)	-0.07	**0.50**	-0.02	0.07	0.16	-0.05	0.02	-0.07	-0.01	0.04
Yo (Ending particle)	0.03	**0.49**	-0.06	0.01	0.21	0.01	-0.02	-0.02	0.25	0.31
Mitai (Auxiliary verb stem)	0.06	**0.49**	0.00	0.01	-0.02	0.00	-0.01	0.16	0.02	-0.03
Ya (Auxiliary verb)	0.04	-0.03	**0.93**	-0.07	0.00	0.01	0.04	0.01	-0.01	-0.04
Hen (Auxiliary verb)	0.03	0.01	**0.64**	-0.05	-0.02	-0.03	-0.01	0.01	-0.03	-0.05
De (Ending particle)	0.01	0.01	**0.59**	-0.03	-0.02	-0.02	0.00	-0.02	-0.01	0.01
Nen (Ending particle)	-0.07	0.02	**0.47**	0.07	0.02	0.04	-0.01	0.01	-0.01	0.04
Haru (Auxiliary verb)	0.00	0.01	**0.45**	0.02	0.02	-0.02	-0.01	-0.02	0.00	-0.02
Masu (Auxiliary verb)	0.22	0.05	-0.03	**0.83**	-0.03	0.04	-0.12	-0.05	0.04	-0.13
Desu (Auxiliary verb)	0.19	0.19	-0.05	**0.71**	-0.05	0.03	-0.08	-0.05	0.28	-0.22
Zu (Auxiliary verb)	0.27	-0.08	0.10	**0.67**	0.01	0.01	0.11	0.07	-0.12	0.09
Wa (Ending particle)	-0.01	-0.07	0.05	-0.02	**0.97**	0.01	-0.01	-0.01	0.12	-0.04
Kashira (Ending particle)	0.03	-0.02	-0.02	0.00	**0.57**	0.02	-0.02	0.04	0.05	-0.12
No (Ending particle)	0.03	0.31	0.01	-0.09	**0.53**	-0.04	0.01	0.03	-0.06	-0.05
Yagaru (Auxiliary verb)	-0.03	0.03	-0.02	0.08	0.02	**0.78**	0.00	-0.01	-0.09	-0.09

Table 5: Continued.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
Ze (Ending particle)	-0.02	0.03	-0.06	0.02	-0.02	**0.74**	-0.03	0.01	-0.08	0.04
I (Ending particle)	0.02	-0.05	0.07	-0.06	0.02	**0.53**	0.05	0.04	0.08	0.15
Ja (Auxiliary verb)	0.05	0.07	-0.03	-0.06	-0.03	0.00	**0.76**	-0.05	-0.02	-0.03
Nou (Ending particle)	0.01	0.02	0.01	-0.05	0.00	0.00	**0.56**	0.00	0.02	-0.04
Ne (Ending particle)	0.02	0.36	-0.03	0.12	0.13	-0.08	0.00	0.02	**0.53**	0.18
Sa (Ending particle)	0.06	0.25	-0.02	-0.17	-0.14	0.05	-0.05	-0.01	0.10	**0.46**

Table 6: Examples of each utterance styles.

Style	Example	Japanese
Neutral	Nani wo shite iru?	何をしている?
Frank	Nani shi teru?	何してる?
Dialect	Nani shi torun ya?	何しとるんや?
Polite	Nani shi te i masu ka?	何していますか?
Feminine	Nani shi te iru no?	何しているの?
Crude	Nani shi te yagaru?	何してやがる?
Aged	Nani shi torun ja?	何しとるんじゃ?
Interrogative	Nani ka shi te nai?	何かしてない?
Approval	Nani ka shi teru no desu ne?	何かしてるのですね?
Dandy	Nani shi te iru no da?	何しているのだ?

(v) Factor 5: This factor mainly included feminine characteristic particles (e.g., wa, kashira, and no) and was referred to as 'feminine style'.

(vi) Factor 6: This factor included relatively crude expressions (e.g., yagaru, and ze). Therefore, it was labelled 'crude style'.

(vii) Factor 7: This factor included expressions indicating aged people (e.g., ja, and nou). Therefore, it was labelled 'aged style'.

(viii) Factor 8: This factor included nai, ka, and mo (loading values of ka and mo are less than 0.4). Since those are related to suspicions, questions, and interrogative forms, it was labelled 'interrogative style'.

(ix) Factor 9: This factor included ne, desu, and yo (loading values of desu and yo are less than 0.4). Those words are used in case of indicating approval or serve as backchannel. Therefore, it was labelled 'approval style'.

(x) Factor 10: This factor included sa, yo, and da (loading values of yo and da are less than 0.4). Those words are used to impart masculinity to a character and also pretentious mood. Therefore, it was labelled 'dandy style'.

Table 6 shows examples of each utterance style. These example sentences signify the same meaning of "What are you doing?" in Japanese. The difference in nuance cannot be expressed in English language in principle.

In order to investigate the relationships between extracted factors and fundamental attributes of fictional characters, average factor scores were calculated for each gender group and each age group (Table 7). In Table 7, bold cells indicate the absolute values exceeding 0.2. Since those tendencies may reflect some prejudice of fiction writers, it is possible that they may not correspond to the real tendencies of Japanese utterances. However, those results would reflect the average tendencies of fictional characters' utterance styles in Japanese novels.

Average Factor 1 (neutral style) scores indicate that aged males were expected to speak traditional Japanese. On the other hand, young female characters were not regarded as the speakers of traditional Japanese. Average Factor 2 (frank style) scores indicate that frank style was widely utilised for female characters. Young male characters used frank styles only exceptionally. The Factor 3 (dialect style) score shows that dialect style was independent of the categories of attributes. Since a dialect style reflects mainly the birthplace of a character, this result seems reasonable. The Factor 4 (polite style) scores were low both in young males and females. This result may correspond to the Factor 2 (frank style) scores. Although it is self-explanatory, Factor 5 score (feminine style) proved feminine utterance style was utilised by female characters. The Factor 6 (crude style) scores were a bit high in young and middle-aged male characters and a bit low in young and middle-aged female characters, and vice versa. The Factor 7 (aged style) scores also proved that aged male characters utilised aged utterance styles. The Factor 8 scores (Interrogative style) indicated that utterance styles about suspicions, questions, and interrogative forms were not dependent on characters' fundamental attributes. It seems reasonable since every character can have suspicions and questions. The Factor 9 scores (approval style) show that aged female frequently used backchannel speech style in fictional texts. On the other hand, young female did not often utilise backchannel speech style. The Factor 10 (dandy style) scores indicate male characters often exhibit masculine mood. Of course, female characters did not utilise those utterance styles.

Those relationships between extracted utterance styles and fundamental attributes (gender and age) of fictional characters revealed that some utterance styles suggest specific categories of characters; therefore, those utterance styles would be utilisable for speaker identification tasks in natural language processing. In addition, those utterance styles would facilitate the generation of more natural dialogues in automatic dialogue generation tasks based on some virtual attributes of speakers. Moreover, those utterance styles may be useful for deducing the personality, social status, and social relationships of speakers and listeners in order to interpret stories in texts.

4. Conclusions

The characteristics of utterances in Japanese novels were analysed by adding several attributes to a randomly extracted

Table 7: Average factor scores for categories about genders and ages.

		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
		Neutral	Frank	Dialect	Polite	Feminine	Crude	Aged	Interrogative	Approval	Dandy
Male	Young	-0.15	**0.26**	0.10	**-0.27**	-0.19	0.15	-0.16	0.08	-0.08	0.16
	Middle	0.06	-0.19	0.01	0.06	**-0.26**	0.08	0.04	-0.01	0.02	0.14
	Aged	**0.25**	-0.18	0.03	0.12	**-0.22**	0.01	**0.60**	0.10	0.10	**0.29**
Female	Young	**-0.22**	**0.47**	-0.06	**-0.23**	**0.38**	-0.15	-0.14	0.03	-0.19	-0.18
	Middle	-0.06	**0.35**	-0.02	-0.05	**0.69**	-0.18	-0.11	0.02	-0.03	**-0.35**
	Aged	0.02	**0.31**	0.05	0.06	**0.22**	-0.05	0.07	0.03	0.17	-0.12

Table 8: BCCWJ title IDs of randomly sampled 100 Japanese novels.

LBa9_00016	LBi9_00004	LBm9_00097	LBq9_00233
LBa9_00076	LBi9_00029	LBm9_00238	LBq9_00236
LBa9_00112	LBi9_00040	LBm9_00244	LBr9_00027
LBb9_00014	LBi9_00176	LBm9_00261	LBr9_00065
LBb9_00028	LBi9_00218	LBm9_00264	LBr9_00081
LBb9_00051	LBj9_00004	LBn9_00018	LBr9_00166
LBb9_00092	LBj9_00127	LBn9_00041	LBr9_00210
LBb9_00147	LBj9_00242	LBn9_00084	LBr9_00232
LBc9_00046	LBj9_00263	LBn9_00235	LBr9_00257
LBc9_00074	LBk9_00127	LBo9_00012	LBs9_00023
LBc9_00086	LBk9_00187	LBo9_00029	LBs9_00173
LBc9_00160	LBk9_00245	LBo9_00060	LBs9_00180
LBd9_00035	LBk9_00269	LBo9_00063	LBs9_00194
LBd9_00066	LBk9_00276	LBo9_00075	LBs9_00247
LBd9_00100	LBl9_00011	LBo9_00240	LBs9_00280
LBd9_00146	LBl9_00102	LBo9_00255	LBt9_00020
LBd9_00185	LBl9_00127	LBp9_00018	LBt9_00059
LBf9_00002	LBl9_00206	LBp9_00034	LBt9_00080
LBf9_00053	LBl9_00210	LBp9_00150	LBt9_00090
LBf9_00132	LBl9_00269	LBp9_00154	LBt9_00100
LBg9_00210	LBm9_00004	LBq9_00019	LBt9_00105
LBh9_00065	LBm9_00027	LBq9_00028	LBt9_00115
LBh9_00082	LBm9_00040	LBq9_00040	LBt9_00134
LBh9_00148	LBm9_00058	LBq9_00083	LBt9_00204
LBh9_00240	LBm9_00068	LBq9_00181	LBt9_00252

Japanese novel corpus, and 5632 annotated utterances (887 data sets) were prepared. Based on the corpus, the characteristics of utterance styles were extracted quantitatively. A chi-square test for particles, auxiliary verbs, and utterance characteristics of genders, and relationship between the speakers and listeners revealed that male utterances included more imperative words, whereas female utterances contained more particle verbs which implied a polite tone. In addition, auxiliary verbs of politeness were more frequently used for ‘coworkers’ and ‘superior authorities’ than for ‘friend’, ‘subordinates’, and ‘enemy’. The results also revealed differences in utterances between close and intimate relationships. Finally, repeated factor analyses revealed seven frequently used utterance styles (neutral, frank, dialect, polite, feminine, crude, aged, interrogative, approval, and dandy). The factor scores indicated relationships between various utterance styles and fundamental attributes of speakers. Thus, results of this study would be utilisable for speaker identification tasks, automatic speech generation tasks, and scientific interpretation of stories and characters.

Although in this study, factor analysis has been done for the large corpus with fundamental attributes, some large corpus with detailed attributes could allow for a more in-depth analysis of the relationship between utterance styles and attributes. Moreover, future research should focus on the validation of usefulness of these results.

Also it would be useful to compare the results with those of similar corpus of other languages [15].

Appendix

See Table 8.

Data Availability

The BCCWJ speaker information corpus is now prepared for release at the National Institute for Japanese Language and Linguistics and it will be available optionally to the users of the BCCWJ.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 18K11991, ‘Analysis for Styles of Utterances and Speaker’s Attributes within Story Texts and Daily Conversation’, and Number 26730168, the NINJAL collaborative research project ‘A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation’, and the NINJAL project ‘Corpus of Everyday Japanese Conversation’.

References

[1] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the Association for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[2] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the Association for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[3] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, "Gender, genre, and writing style in formal written texts," *Text - Interdisciplinary Journal for the Study of Discourse*, vol. 23, no. 3, pp. 321–346, 2006.

[4] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern et al., "Personality, gender, and age in the language of social media: the open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, Article ID e73791, 2013.

[5] S. Goswami, S. Sarkar, and M. Rustagi, "Stylometric Analysis of Bloggers' Age and Gender," in *Proceedings of the Third International ICWSM Conference*, pp. 214–217, 2009.

[6] Kinsui. Satoshi, *Virtual Japanese: Mystery of Functional Words , Iwanami Shoten*, In Japanese, Iwanami Shoten, Tokyo, 2003.

[7] S. Okamoto, "Social context, linguistic ideology, and indexical expressions in Japanese," *Journal of Pragmatics*, vol. 28, no. 6, pp. 795–817, 1997.

[8] S. Okamoto, "Situated Politeness: Manipulating Honorific and Non-Honorific Expressions in Japanese Conversations," *Pragmatics*, vol. 9, no. 1, pp. 51–74, 1999.

[9] H. Murai, "Factor Analysis of Japanese Daily Utterance Styles," in *LREC 2018 Joint Workshop LB-ILR2018 and MMC2018 Proceedings*, pp. 26–29, 2018.

[10] C. Miyazaki, T. Hirano, R. Higashinaka et al., "Fundamental analysis of linguistic expression that contributes to characteristics of speaker," in *Proceedings of the Association for Natural Language Processing*, pp. 232–235, 2014.

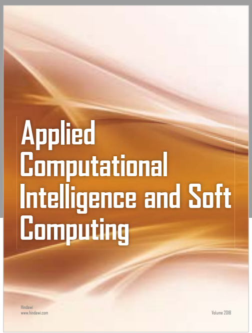
[11] R. Shen, K. Hideaki, K. Ohta, and M. Takeshi, "Towards the Text-level Characterization Based on Speech Generation," *Journal of Information Processing Society of Japan*, vol. 53, no. 4, pp. 1269–1276, 2012.

[12] H. Murai, "Towards agent estimation system for story text based on agent vocabulary dictionary," in *IPSJ Symposium Series*, vol. 2016, pp. 209–214, 2016.

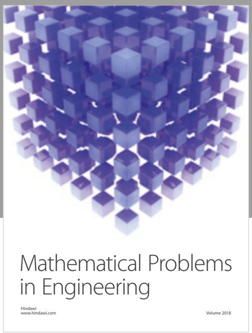
[13] K. Maekawa, M. Yamazaki, T. Ogiso et al., "Balanced corpus of contemporary written Japanese," *Language Resources and Evaluation*, vol. 48, no. 2, pp. 345–371, 2014.

[14] S. Ogawa, "Gender difference of spoken language," *The Society for Gender Studies in Japanese*, no. 4, pp. 26–39, 2004.

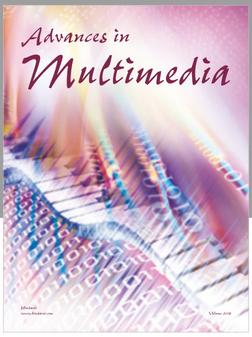
[15] B. Verhoeven and W. Daelemans, "CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text," in *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 3081–3085, Iceland, May 2014.

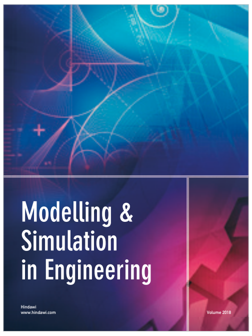

Applied Computational Intelligence and Soft Computing

The Scientific World Journal


Mathematical Problems in Engineering

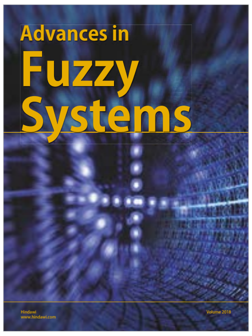
Journal of Engineering

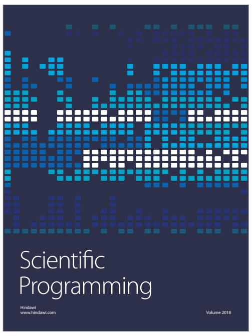

Advances in Multimedia

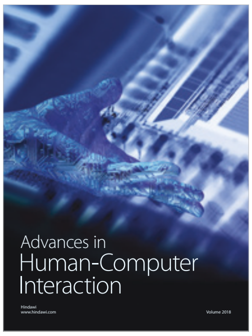

Modelling & Simulation in Engineering

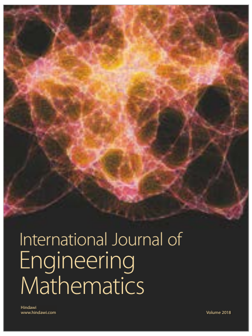
Advances in Artificial Intelligence

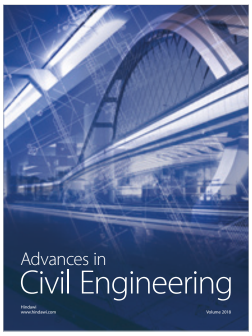
International Journal of Reconfigurable Computing

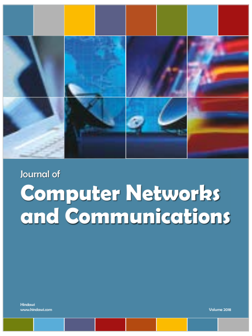

Advances in Fuzzy Systems


Scientific Programming


Advances in Human-Computer Interaction

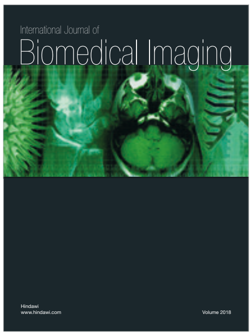

International Journal of Engineering Mathematics


Advances in Civil Engineering

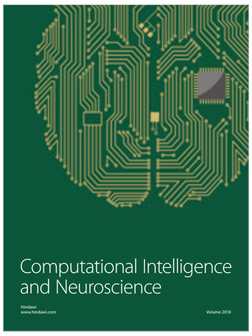

Journal of Computer Networks and Communications

Journal of Robotics

International Journal of Computer Games Technology


International Journal of Biomedical Imaging

Journal of Electrical and Computer Engineering


Computational Intelligence and Neuroscience