

Japanese Daily Utterance Styles: A Factor Analysis based on Balanced Corpus

Hajime Murai

Future University Hakodate
116-2 Kamedanakano-cho, Hakodate,
Hokkaido, Japan
h_murai@fun.ac.jp

Abstract

It is often considered more difficult to extract fundamental utterance styles in everyday conversation than in fictional utterances. This is because the characteristics of utterance styles are exaggerated in fictional utterances. However, by referring to a large-scale corpus of daily conversations, it is possible to identify the fundamental patterns of everyday Japanese utterance styles. This study employs the statistical method of factor analysis to identify the characteristics of utterance styles within the Corpus of Everyday Japanese Conversation - a gender and age balanced corpus. Eight factors ("Neutral style," "Dialect style," "Frank style," "Polite style," "Feminine style," "Crude style," "Series style," and "Parallel style") were extracted quantitatively. The results suggest that "Series style" and "Parallel style" are unique to everyday conversation. On the other hand, "Aged style," "Interrogative style," "Approval style," and "Dandy style" (found in utterances in written fiction) were not found. Unlike previous studies, these results are based on a balanced corpus.

1 Introduction

Utterance styles are affected by various factors, such as gender, age, context, cultural setting, social background, personalities of the characters, and the mood of the moment.

Elsewhere in the literature, characterized written styles is applied to texts of various kinds for text

categorization and author identification tasks (Zheng et al., 2006; Stamatatos, 2009). Previous research has analyzed the characteristics of utterance styles mainly on the basis of gender, or age (Argamon et al., 2006; Schwartz et al., 2013; Goswami et al., 2009).

In the case of Japanese fictional utterances (in novels or essays), an important way characters can be differentiated is on the basis of utterance style. This popular technique helps readers understand each character's personality (Kinsui, 2003). These utterance styles can be detected by comparing the frequency of function words in utterances. Furthermore, fundamental patterns of utterance styles composed of particles and auxiliary verbs can be identified by conducting a factor analysis of a fictional corpus (Murai, 2018A).

In the field of Japanese everyday conversation, the main research topics tended to focus on grammatical characteristics and pragmatic semantics (Seto and Kishi, 2015), as well as the relationships between single attributes (such as politeness and gender) and utterance styles (Kurosawa, 2010). It is considered more difficult to extract fundamental utterance styles in real, daily conversation than in fiction writing, because distinct utterance styles are often exaggerated in conversations between fictional characters (particularly in entertainment contents). Therefore, explorations of Japanese utterance styles in daily conversation have tended to employ case studies (Miyazaki et al., 2014) and psychological experimental approaches (Shen et al., 2012).

Previous attempts to extract utterance styles in daily conversation (Murai, 2018B) have yielded mixed results because of an unbalanced corpus. A total pattern of Japanese utterance styles identified using a quantitative analysis of a real-life, balanced corpus has so far been lacking. However, by drawing on a large-scale, balanced corpus of daily conversation – in this case, the trial version of the new Corpus of Everyday Japanese Conversation (Koiso et. al., 2016) – it has been possible to identify fundamental patterns using the statistical method of factor analysis. In addition, the present study examined the difference between utterance styles in fiction writing and in daily conversation. The difference of utterance styles had been predicted by previous linguistic studies (Kinsui, 2003) but it had not been examined quantitatively.

2 Corpus used in Analysis

The trial version of the Corpus of Everyday Japanese Conversation (CEJC) is composed of 126 real-life Japanese dialogues. The duration of these recordings is 3015 minutes. A total of 463 speakers were recorded their real-life natural dialogues (some speakers were included in several dialogues). The speakers are native Japanese speakers of various ages. The dialogues take place in various situations, such as home, school, workplace, restaurant and leisure. In addition, speakers have different relationships with each other, such as family, friends, co-workers, teachers, and students.

The factor analysis grouped utterances by speaker in 126 conversation scenes. In total, 435 utterance sets were identified in the CEJC, (excluding very reticent speakers who spoke less than ten words, for statistical reasons).

The attributes of the speakers in the 435 utterance sets are given in Table 1.

Age	Male	Female	Total
10 to 29	51	40	91
30 to 49	54	92	146
50 to 69	48	73	121
Over 70	19	18	37
Total	172	223	395

Table 1: Speaker details for utterance sets in the CEJC

Table 1 does not include 40 speakers whose attributes are unknown. These are accidental

participants such as restaurant employees. In the analysis, these unknown speakers were included as the 435 utterance data set, except for very reticent speakers as mentioned above.

3 Characteristics of Utterance Styles

In this study, the frequency with which function words occur in utterances was adopted as a characteristic of utterance style, because in many Japanese novels, different usage patterns of function words are used to express characters' personalities (Kinsui, 2003). In the Japanese language, function words usually correspond to particles and auxiliary verbs. Therefore, the statistical significance of the frequency with which particles and auxiliary verbs occur was analyzed using factor analysis (Murai, 2018A) for a fictional corpus. The CEJC provides morphologically analyzed data sets for the conversation texts; therefore, particles and auxiliary verbs in utterances could be extracted from the 435 data set units and counted. The frequency with which the top 30 particles and auxiliary verbs appeared is shown in Table 2.

4 Factor Analysis for Utterance Styles

4.1 Factor Analysis for Daily Conversation

To extract the typical utterance styles of Japanese daily conversation, a factor analysis was carried out to establish the frequency with which particles and auxiliary verbs were used. Owing to statistical limitations, 83 function words (where the frequency of those particles and auxiliary verbs exceeded 10) were selected, and 83 dimensional word frequency vectors were extracted for each speaker, in each scene. The Promax rotation method was used and a parallel analysis performed to determine the number of factors involved. After the factor analysis, less significant words (with a maximum factor loading of < 0.3) were eliminated and a factor analysis was repeatedly performed for the eliminated data set. Finally, after performing the factor analysis four times, eight factors were identified. The resultant factor loadings are shown in Table 3. The bold font signifies cells where the absolute value of factor scores exceeded 0.3.

Table 4 depicts average factor scores for each age / gender category in the CEJC (as in Table 1). Table 5 depicts another average factor score for

each situational category. “Speakers” signifies the number of speakers who participated in each dialogue of that category. In both tables 4 and 5, bold font designates cells whose absolute value of factor scores exceeded 0.2.

Word and part of speech	Frequency
Auxiliary verb "Da"	29708
Final particle "Ne"	17250
Auxiliary verb "Ta"	14493
Connective particle "Te"	13466
Quasi-particle "No"	13335
Case particle "No"	10080
Case particle "De"	9852
Auxiliary particle "Ka"	9689
Case particle "Ga"	9496
Incidental particle "Mo"	9318
Incidental particle "Wa"	9112
Final particle "Yo"	8954
Case particle "Ni"	8870
Case particle "To"	8381
Auxiliary verb "Teru"	7947
Auxiliary particle "Tte"	7316
Auxiliary verb "Nai"	6711
Auxiliary verb "Desu"	6424
Final particle "Ka"	5919
Connective particle "Kara"	5784
Connective particle "Keredo"	4118
Final particle "No"	3756
Final particle "Na"	3589
Final particle "Sa"	3371
Case particle "Wo"	3253
Auxiliary verb "Masu"	2881
Auxiliary verb "Chau"	1957
Connective particle "To"	1794
Auxiliary verb "Tuu"	1652
Case particle "Kara"	1482

Table 2: Top 30 frequently appearing words

Eight factors corresponded with utterance patterns that frequently appeared in daily conversation in the CEJC. The characteristics of each, as well as an explanation of how each was named, are provided below:

Factor 1: Included the most frequently used general function particles and auxiliary verbs, such

as the case particles “Wo,” “To,” “Ni,” “Ga,” “No,” “De,” and “Kara”. However, Factor 1 did not include words that indicated specific attributes. In other words, it represented a ‘Neutral style’ of utterance.

This utterance style seems to be commonly used by middle-aged females (Table 4). It is also generally employed in situations that are not particularly intimate, such as school, business or service situations, like shopping (Table 5).

Factor 2: Included particles and auxiliary verbs such as “Hen,” “Yan,” “De,” and “Nen”. These words are frequently used in Japanese dialects such as Kansai-ben. Therefore, Factor 2 is referred to as “Dialect style”.

This utterance style is often used by middle-aged males (Table 4). The reason would be that the category of middle aged men includes more people from Kansai region than other categories.

Factor 3: Included final particles such as “Jan,” “Yo,” “Mono,” “Ke,” and “Sa.” These are characteristic of informal, frank communication styles. Therefore, Factor 3 is referred to as “Frank style.”

Table 4 shows that this factor is strongly associated with males between the ages of 10 and 29. In addition, Table 5 shows that this utterance style is frequently used in family relationships and service situations. This suggests, for instance, that customers often use frank utterance styles when speaking to sales clerks in shopping situations.

Factor 4: Included the auxiliary verbs “Desu” and “Masu.” These are clearly related to Japanese honorific utterance styles. Therefore, Factor 4 is referred to as “Polite style.”

This style is generally used in less intimate situations such as schools or businesses, or in service settings like shopping (Table 5). It is similar to Factor 1 in this respect. This factor is common amongst young males (see Table 4), it is also often used in dialogues at school (which could also include young male students in Table 5).

Factor 5: Included feminine characteristic particles (e.g. “Wa,” “Kashira,” and “No”) and is thus referred to as “Feminine style”. This utterance style is related to middle-aged and elderly females as expected (in Table 4). However, young women do not appear to use this traditional utterance style.

	F1	F2	F3	F4	F5	F6	F7	F8
Case particle "Wo"	1.04	-0.01	-0.12	0.04	-0.09	0.00	0.10	-0.26
Case particle "To"	0.99	-0.01	-0.19	-0.02	0.00	-0.04	-0.12	0.31
Auxiliary particle "Ka"	0.92	-0.05	-0.13	-0.06	0.06	0.00	-0.33	0.36
Auxiliary particle "Tari"	0.90	0.01	-0.32	0.00	-0.06	-0.02	-0.15	0.11
Case particle "Ni"	0.85	-0.02	-0.09	0.00	0.09	0.07	0.12	0.07
Case particle "Ga"	0.84	-0.02	0.06	0.09	-0.01	0.03	0.11	-0.07
Incidental particle "Mo"	0.80	-0.01	0.01	0.01	0.13	-0.04	-0.08	0.18
Connective particle "To"	0.80	-0.03	-0.18	0.08	0.09	0.12	-0.06	0.02
Connective particle "Keredo"	0.79	-0.04	0.05	0.01	0.02	-0.02	-0.10	0.21
Incidental particle "Wa"	0.78	-0.01	0.09	0.09	0.01	0.00	0.03	-0.05
Connective particle "Te"	0.78	0.03	-0.13	0.03	0.05	-0.02	0.26	0.16
Case particle "No"	0.78	0.01	0.13	0.05	-0.09	-0.01	0.16	-0.05
Auxiliary particle "Tte"	0.76	0.01	0.25	-0.04	-0.10	-0.02	0.01	0.08
Auxiliary verb "Da"	0.70	-0.07	0.34	-0.08	0.20	0.05	-0.13	-0.02
Case particle "De"	0.68	0.00	0.10	0.09	0.11	0.00	0.08	0.05
Quasi-particle "No"	0.64	-0.04	0.16	0.29	0.07	-0.01	-0.04	0.02
Final particle "Na"	0.62	0.29	0.14	0.07	0.12	0.00	-0.12	-0.10
Auxiliary verb "Teru"	0.59	0.05	0.31	-0.08	0.03	-0.01	0.04	0.16
Auxiliary verb "Seru"	0.57	-0.03	0.00	-0.11	-0.05	0.02	-0.05	-0.15
Auxiliary verb "Ta"	0.55	0.01	0.18	-0.03	0.17	0.05	0.09	0.14
Final particle "Ne"	0.54	-0.09	0.15	0.09	0.53	-0.09	-0.17	-0.20
Auxiliary verb "Rareru"	0.54	-0.01	0.19	-0.07	-0.22	0.01	0.09	-0.02
Connective particle "Kara"	0.54	0.04	0.26	-0.13	0.06	-0.01	0.16	0.15
Connective particle "Nagara"	0.52	0.01	0.10	0.06	-0.05	-0.04	-0.09	-0.04
Auxiliary verb "Reru"	0.51	0.03	0.31	-0.08	-0.20	0.02	0.04	0.13
Connective particle "Shi"	0.50	0.01	0.08	-0.02	0.00	0.04	-0.12	0.33
Case particle "Kara"	0.48	-0.02	0.02	-0.03	0.08	-0.02	0.33	0.08
Auxiliary verb "Nai"	0.47	-0.04	0.34	0.03	0.03	0.00	0.10	0.14
Auxiliary particle "Kurai"	0.43	0.01	-0.10	0.22	0.12	-0.01	0.09	0.20
Connective particle "Ba"	0.41	-0.02	0.04	0.23	-0.14	0.01	0.27	0.08
Auxiliary particle "Dake"	0.40	0.08	0.12	0.12	0.02	-0.08	0.10	0.06
Auxiliary verb "Hen"	0.01	1.00	0.00	-0.02	0.00	-0.14	0.03	0.02
Final particle "Yan"	0.01	0.96	0.01	-0.02	-0.01	0.06	-0.02	0.02
Final particle "De"	-0.02	0.93	0.00	0.01	0.03	-0.16	0.08	-0.01
Final particle "Nen"	-0.04	0.91	-0.07	0.00	0.04	0.27	-0.02	-0.01
Final particle "Jan"	0.15	-0.03	0.62	-0.27	-0.19	0.07	0.07	0.17
Final particle "Yo"	0.15	-0.01	0.59	0.19	0.31	-0.01	0.05	-0.18
Final particle "Mono"	-0.08	0.03	0.55	0.15	0.14	0.05	0.00	0.04
Auxiliary verb "Tuu"	0.34	-0.01	0.48	0.10	-0.04	0.05	0.08	-0.17
Final particle "Ke"	0.12	0.02	0.47	0.08	0.04	0.05	-0.24	0.01
Final particle "Sa"	0.42	-0.04	0.45	-0.32	-0.01	-0.04	-0.02	0.01
Auxiliary particle "Sura"	0.04	-0.01	0.40	0.06	-0.09	-0.04	-0.01	-0.11

Table 3-1: Results of factor analysis of frequently appearing function words in the CEJC

	F1	F2	F3	F4	F5	F6	F7	F8
Auxiliary verb "Desu"	-0.03	0.01	-0.02	1.11	-0.28	0.00	0.04	0.02
Auxiliary verb "Masu"	0.28	-0.01	-0.28	0.80	-0.10	0.02	0.07	-0.10
Final particle "Ka"	0.34	-0.03	0.12	0.61	0.03	-0.01	-0.05	-0.05
Connective particle "Tutu"	0.00	-0.01	0.20	0.34	-0.03	-0.04	-0.11	-0.05
Final particle "Kashira"	-0.02	-0.06	-0.12	-0.06	0.59	-0.02	0.07	-0.17
Final particle "Wa"	0.00	0.10	0.01	-0.14	0.57	0.00	-0.01	-0.09
Final particle "No"	0.20	-0.01	0.38	-0.33	0.39	-0.03	0.06	0.04
Auxiliary verb "Chau"	0.26	-0.09	0.13	0.05	0.36	0.01	0.13	0.00
Auxiliary verb "Yagaru"	-0.13	-0.02	-0.01	0.04	-0.04	0.76	-0.08	0.08
Auxiliary verb "Beshi"	-0.05	-0.03	-0.04	-0.04	0.08	0.63	-0.07	0.00
Connective particle "Ga"	0.30	0.05	-0.09	-0.01	0.04	0.49	-0.04	-0.18
Final particle "Zo"	0.00	0.00	0.17	-0.04	-0.13	0.34	0.04	-0.05
Case particle "He"	0.03	0.03	-0.06	0.01	0.06	-0.06	0.50	-0.11
Connective particle "Tatte"	-0.04	0.00	0.01	-0.04	0.00	-0.02	0.40	0.03
Auxiliary verb "Teku"	0.33	-0.01	0.01	0.00	0.02	-0.10	0.35	0.03
Auxiliary verb "Toku"	0.00	-0.01	-0.09	0.11	0.22	0.26	0.34	0.05
Auxiliary verb "Saseru"	0.24	0.00	-0.07	-0.07	-0.15	-0.03	-0.01	0.37
Auxiliary verb "Tai"	0.31	0.02	0.07	0.19	-0.07	-0.04	-0.08	0.34
Auxiliary particle "Shika"	0.14	0.01	0.25	0.12	-0.13	0.11	0.06	0.32

Table 3-2: Results of factor analysis of function words frequently appearing in the CEJC

	Age	F1	F2	F3	F4	F5	F6	F7	F8
Male	10 to 29	0.09	0.00	0.42	0.22	-0.23	0.43	-0.12	0.19
	30 to 49	0.12	0.39	0.06	0.11	-0.09	0.10	-0.15	-0.08
	50 to 69	-0.18	-0.09	-0.13	-0.16	-0.23	-0.11	-0.15	-0.30
	Over 70	-0.04	-0.07	-0.12	-0.16	-0.03	-0.07	0.69	-0.11
Female	10 to 29	-0.21	-0.06	-0.17	-0.17	-0.32	-0.08	-0.12	-0.01
	30 to 49	0.26	-0.08	0.19	0.14	0.45	-0.02	0.07	0.27
	50 to 69	0.18	-0.01	0.08	0.11	0.29	-0.05	0.05	0.16
	Over 70	-0.26	-0.09	-0.27	-0.14	0.18	-0.12	0.54	-0.27

Table 4: Average factor scores for each gender / age category in the CEJC

	Speakers	F1	F2	F3	F4	F5	F6	F7	F8
Family	117	0.00	0.04	0.24	-0.28	-0.09	0.22	0.21	0.08
Family and relatives	58	-0.32	-0.09	-0.22	-0.15	-0.02	-0.16	0.20	-0.14
Friends	156	0.07	0.05	0.08	0.11	0.15	-0.02	-0.09	0.11
Teachers and students	6	1.03	-0.07	-0.03	1.83	0.15	0.01	-0.54	0.20
Business relationship	14	0.36	-0.09	-0.27	0.52	-0.01	-0.07	0.19	-0.25
Co-worker	22	-0.20	-0.07	-0.09	-0.05	0.00	-0.19	-0.43	-0.06
Sales clerk and customer	7	0.28	-0.09	0.26	0.26	0.20	-0.04	-0.03	0.29

Table 5: Average factor scores for each situation category in the in the CEJC

Factor 6: Included relatively crude expressions such as “Yagaru,” and “Zo” and connective particle “Ga” which also has a crude nuance. Therefore, it is labelled “Crude style.”

This utterance style is common amongst young males (Table 4). It is not suitable in formal situations; and is therefore only associated with the family situation (Table 5).

Factor 7: Included the case particle “Kara,” “He,” and connective particle “Tatte”. “Kara” and “Tatte” are often used to signify logical connections, such as cause and effect. It is therefore associated with a series of connected utterances using particles that represent logical relationships. And in consequence is referred to as “Series style.” This style is used mainly by elderly males and females (Table 4).

Factor 8: Included case particle “To,” auxiliary particle “Ka,” and connective particle “Shi”. Both particles are used to juxtapose sentences or phrases, much like the English words “and” or “or”. In contrast with Factor 7, this utterance style indicates that utterances are connected in a parallel fashion using particles for juxtaposition. Therefore, it is labelled “Parallel style.” This utterance style is most often used by young males and middle age females (Table 4).

Those eight factors and factor scores seem to reflect daily use of utterance styles as mentioned in explanations of factors. However, in some case, detailed differences were not discriminated. For instances, utterance styles in shops include both sales clerk and customer (in Table 5), and the polite utterance style of clerks and the frank utterance style of customers would be combined in the result. Because the utterances were tagged according to situation.

4.2 Comparison with Factor Analysis for Fiction Writings

In order to establish which characteristics might be unique to utterance styles in real daily conversation, the results of the factor analysis were compared with similar results for utterance styles within fiction writing (Murai, 2018A).

The analysis of utterance styles in novels are based on a random sampling of dialogues in Japanese novels. These dialogues are drawn from a subset of texts within the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). These texts from Japanese

novels, are included under the Nippon Decimal Classification number 913. One hundred texts were randomly sampled from Japanese novels included in the library-based corpus in the BCCWJ. In addition, a speaker database provides gender and age attributes for each speaker who appears in the novel texts. Each data entry used in the factor analysis consists of a 100-dimensional vector for all utterances attributed to a single fictional character. Those dimensions indicate the frequency with which 100 types of commonly used particles and auxiliary verbs are used. Because of statistical limitations, 7576 utterance data sets (utterance sets of 7576 speakers in fiction writings) with total frequencies higher than 20 were selected from 11860 data sets.

As a result, ten factors were identified in fiction writing utterance data: “Neutral style,” “Frank style,” “Dialect style,” “Polite style,” “Feminine style,” “Crude style,” “Aged style,” “Interrogative style,” “Approval style,” and “Dandy style” (Murai, 2018A).

The first six factors (“Neutral style,” “Frank style,” “Dialect style,” “Polite style,” “Feminine style,” and “Crude style”) are nearly as common in the utterance styles of daily conversation between real people, and the utterance styles used in fictional conversation between fictional characters. These six utterance styles are therefore clearly characteristic of Japanese speech in general. Examples are shown in Table 6. The sentences used in the example all have the same meaning, “What are you doing?” in Japanese. The difference in nuance cannot be expressed in the English language.

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 ten nen?	何してんねん？
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？

Table 6: Examples of six common utterance styles

Figure 1 depicts the relationships between daily conversation and fiction writing utterances through common particles and auxiliary verbs. The left side of the figure shows factors for daily conversation utterances, and right side shows factors for utterances in fiction writing. Common particles and auxiliary verbs are shown in the middle with lines connecting them to related factors.

In Figure 1, most of the factors have some words in common with a factor from the other side. However, “Aged style” in fictional writing has no common words with other factors. “Aged style” includes the auxiliary verb “Ja” and final particle “Nou.” This style is often used when representing

aged people in Japanese fiction, but it would be unrealistic utterance style to use in real life.

On the other hand, some words of “Serial style” and “Parallel style” used in daily conversation are combined in “Neutral style” in fiction writing. Moreover, some words that are included in “Frank style” in fiction writing are combined in “Neutral style” and “Feminine style” in daily conversation. Therefore, in daily conversation, “Neutral style” tends to be more “frank” (or forthright) than in fiction writing. This result may be influenced by the fact that the CEJC includes more family and friend situations than other situations (Table 5).

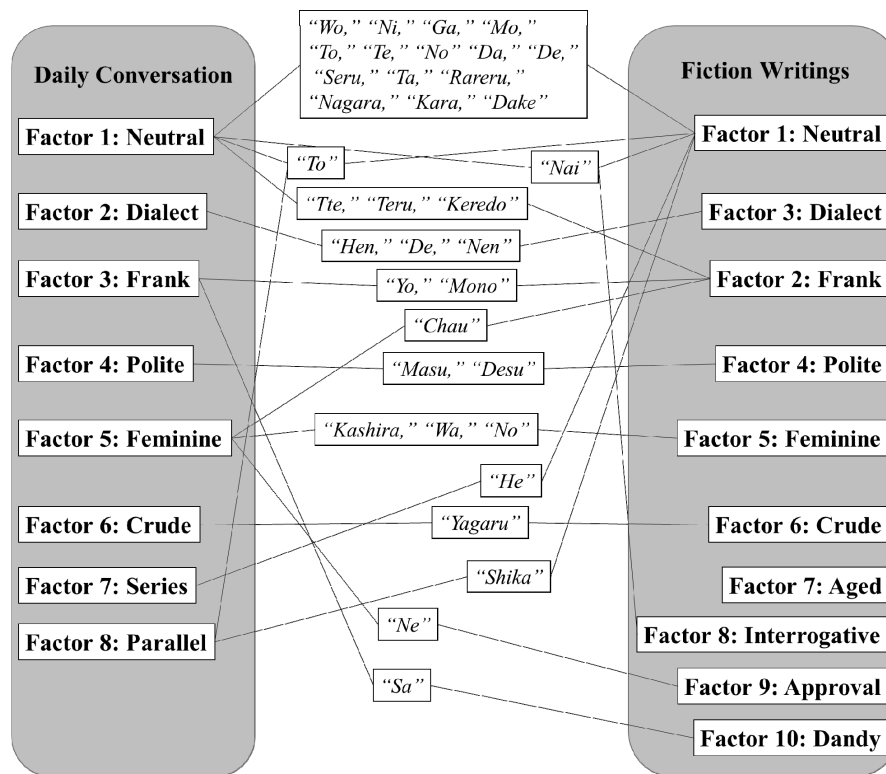


Figure 1: Relationship between daily conversation styles (CEJC) and fiction writing styles (BCCWJ)

5 Conclusion and Future Work

Japanese utterance styles in daily conversation were identified on the basis of a factor analysis for particles and auxiliary verbs in the CEJC. The relationships between these factors on the one hand and the speakers’ attributes (ages and genders) and

conversation settings on the other, were analyzed. As a result, eight factors were extracted: “Neutral style,” “Dialect style,” “Frank style,” “Polite style,” “Feminine style,” “Crude style,” “Series style,” and “Parallel style”.

In addition, a comparison with the utterance styles of characters in fiction writings was done. The results show that six factors are almost

identical between real daily conversation and fiction writings. However, “Neutral style” in daily conversation is more “frank” than in fiction writings. In addition, “Aged style” in fiction writing is a distinctly fictionalized or imaginary style that has no real-life use.

Because the utterances were tagged according to situation, some detailed characteristics of utterance styles were combined in the result of this research (in Table 5). If detailed relationships between speaker and listener for each utterance (e.g. parents to children, superior to subordinate, sales clerk to customer, or teachers to students) were to be added to the CEJC, more detailed utterance styles could be extracted.

Based on six fundamental utterance styles, it would be capable to generate more natural utterances automatically by utilizing natural language processing techniques in the future.

Acknowledgments

This work was supported by JSPS KAKENHI Grant 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

Akiko Kurosawa. 2010. The sentence-final forms used in Meidai Dialogue Corpus: Does the plain style differ from the polite style? *Yamagata University Working Papers in International Education*, 2:3–11. (In Japanese)

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, Yoshiro Matsuo, and Satoshi Sato. 2014. Fundamental Analysis of Linguistic Expression that Contributes to Characteristics of Speaker. In the Proceedings of the Association for Natural Language Processing, pp. 232–235. (In Japanese)

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the Association for Information Science and Technology*, 60(3): 538-556.

Hajime Murai. 2018A. Factor Analysis of Utterances in Japanese Fiction-writing Based on BCCWJ Speaker Information Corpus. *Advances in Human-Computer Interaction*, vol. 2018, Article ID 5056268, 9 pages.

Hajime Murai. 2018B. Factor Analysis of Japanese Daily Utterance Styles. *LREC 2018 Joint Workshop LB-ILR2018 and MMC2018 Proceedings*, 26-29.

Hanae Koiso, Tomoyuki Tsuchiya, Ryoko Watanabe, Daisuke Yokomori, Masao Aizawa, and Yasuharu Den. 2016. Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation. *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, 4434-4439.

Hansen Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9): 1-16.

Kazuma Seto, and Yoshiki Kishi. 2015. Construction of a Dialogue System Using a Speech Type of Estimation by Adjacency. *Proceedings of Information Processing Society of Japan 2015*, 131–132. (In Japanese)

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(22): 345–371.

Raymond Shen, Hideaki Kikuchi, Katsumi Ohta, and Takeshi Mitamura. 2012. Towards the text-level characterization based on speech generation. *Journal of Information Processing Society of Japan*, 53(4):1269–1276. (In Japanese)

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques. *Journal of the Association for Information Science and Technology*, 57(3): 378-393.

Satoshi Kinsui. 2003. *Virtual Japanese: Mystery of Functional Words*. Iwanami Shoten, Tokyo. (In Japanese)

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2006. Gender, Genre, and Writing Style in Formal Written Texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3): 321-346.

Sunit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric Analysis of Bloggers' Age and Gender. *Proceedings of the Third International ICWSM Conference*, 214-217.