Factor Analysis of Japanese Daily Utterance Styles

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

Abstract

It may be more difficult to extract fundamental utterance styles in real-life daily conversation than those in fictional utterances because the characteristics of utterance styles are exaggerated in fictional utterances. However, by utilizing a large-scale corpus of daily conversation, it is possible to identify the fundamental patterns of Japanese utterance styles. In this study, the NUCC was targeted and extraction of the characteristics of utterance styles was carried out using the statistical method of factor analysis. As a result, five factors ("Average style in NUCC," "Avoid affirmation style," "Frank teenager style," "Dialect style," and "Polite style") were extracted quantitatively. Compared to fictional utterance styles, "Avoid affirmation style" is unique in real daily conversation. On the other hand, "Crude style" and "Hearsay style" do not appear. Although the similarities between the fictional corpus and the NUCC support the validity of the result, the factors were impacted by bias in the corpus. It would be desirable to utilize a speaker-balanced daily conversation corpus for a more precise analysis.

Keywords: Utterance, Style, Japanese

1. Introduction

Utterance styles are affected by various attributes, such as gender, age, situation, cultural settings, social backgrounds, personalities of the characters, and the mood of a scene. In the case of Japanese fictional utterances (in novels or general story texts), each character is differentiated based on their utterance styles; it is a popular technique used to help readers understand each character's personality (Kinsui, 2003). These utterance styles can be detected by comparing frequencies of function words in utterances (Murai, 2017A). Moreover, fundamental patterns of utterance styles can be extracted by a factor analysis of a fictional corpus (Murai, 2017B).

In the field of Japanese real conversation in daily life, the main research topics have been general grammatical characteristics, pragmatic semantics (Seto, 2015), and the relationships between specific single attributes (such as politeness and gender) and utterance styles (Kurosawa, 2010). A fundamental total pattern of Japanese utterance styles has not been examined quantitatively based on a real corpus. It may be more difficult to extract fundamental utterance styles in real-life daily conversation because the distinct utterance styles may tend to be exaggerated in conversations between fictional characters (particularly in entertainment content). Therefore, case study approaches and psychological experimental approaches have been used in the field of Japanese utterance styles of daily conversation (Miyazaki, 2014; Shen, 2012).

However, by utilizing a large-scale corpus of daily conversation, it is possible to identify the fundamental patterns of Japanese utterance styles. In this study, the Nagoya University Conversation Corpus (NUCC) was targeted and extraction of the characteristics of utterance styles was carried out using the statistical method of factor analysis.

2. Corpus for Utterance Analysis

The NUCC is composed of transcriptions of 129 uncontrolled, natural conversations between or among friends, family members, or colleagues. Each conversation has two to four participants and lasts 30 to 60 minutes. The participants are 198 native Japanese speakers of various ages and from diverse academic backgrounds (Fujimura, 2012). For the factor analysis, utterances were grouped by each speaker in 129 conversation scenes. In total, 296 utterance sets were obtained from the NUCC (excluding one very reticent speaker for statistical reasons). The attributes of the speakers of the 296 utterance sets are given in Table 1.

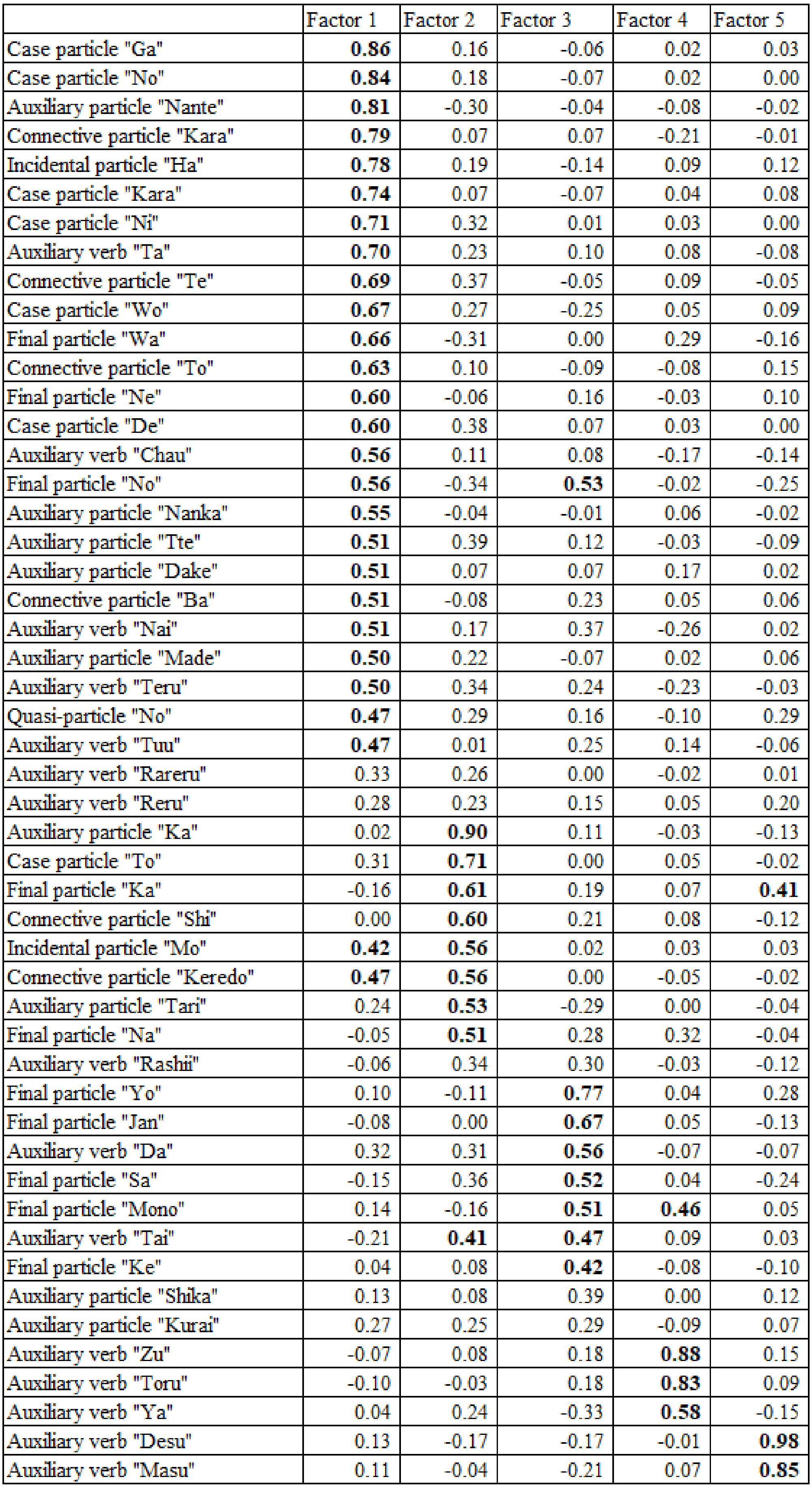
	Female	Male	Total
10s	15	2	17
20s	116	26	142
30s	43	1	44
40s	21	8	29
50s	22	4	26
60s	26	4	30
Over 70s	7	0	7
Unknown	1	0	1
Total	251	45	296

Table 1: Speaker details for utterance sets in the NUCC

It clear that the gender and age balance of the NUCC is biased. However, it is the only large-scale everyday conversation Japanese corpus available. From this table, it is expected that the characteristics of the utterances of young women will be prominently featured.

3. Characteristics in Utterance Styles

In this study, the frequencies of function words in utterances were adopted as characteristics of text style because in many Japanese novels, different usage patterns of function words are used to exhibit characters' personalities (Kinsui, 2003). In the Japanese language, function words mainly correspond to particles and auxiliary verbs. Therefore, the statistical significances of the frequencies of particles and auxiliary verbs were analyzed using factor analysis (Murai, 2017B). The NUCC provides morphologically analyzed data sets for the included conversation texts. Therefore, particles and auxiliary verbs in utterances were extracted and counted from the 296 data set units.



	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Case particle "Ga"	**0.86**	0.16	-0.06	0.02	0.03
Case particle "No"	**0.84**	0.18	-0.07	0.02	0.00
Auxiliary particle "Nante"	**0.81**	-0.30	-0.04	-0.08	-0.02
Connective particle "Kara"	**0.79**	0.07	0.07	-0.21	-0.01
Incidental particle "Ha"	**0.78**	0.19	-0.14	0.09	0.12
Case particle "Kara"	**0.74**	0.07	-0.07	0.04	0.08
Case particle "Ni"	**0.71**	0.32	0.01	0.03	0.00
Auxiliary verb "Ta"	**0.70**	0.23	0.10	0.08	-0.08
Connective particle "Te"	**0.69**	0.37	-0.05	0.09	-0.05
Case particle "Wo"	**0.67**	0.27	-0.25	0.05	0.09
Final particle "Wa"	**0.66**	-0.31	0.00	0.29	-0.16
Connective particle "To"	**0.63**	0.10	-0.09	-0.08	0.15
Final particle "Ne"	**0.60**	-0.06	0.16	-0.03	0.10
Case particle "De"	**0.60**	0.38	0.07	0.03	0.00
Auxiliary verb "Chau"	**0.56**	0.11	0.08	-0.17	-0.14
Final particle "No"	**0.56**	-0.34	**0.53**	-0.02	-0.25
Auxiliary particle "Nanka"	**0.55**	-0.04	-0.01	0.06	-0.02
Auxiliary particle "Tte"	**0.51**	0.39	0.12	-0.03	-0.09
Auxiliary particle "Dake"	**0.51**	0.07	0.07	0.17	0.02
Connective particle "Ba"	**0.51**	-0.08	0.23	0.05	0.06
Auxiliary verb "Nai"	**0.51**	0.17	0.37	-0.26	0.02
Auxiliary particle "Made"	**0.50**	0.22	-0.07	0.02	0.06
Auxiliary verb "Teru"	**0.50**	0.34	0.24	-0.23	-0.03
Quasi-particle "No"	**0.47**	0.29	0.16	-0.10	0.29
Auxiliary verb "Tuu"	**0.47**	0.01	0.25	0.14	-0.06
Auxiliary verb "Rareru"	0.33	0.26	0.00	-0.02	0.01
Auxiliary verb "Reru"	0.28	0.23	0.15	0.05	0.20
Auxiliary particle "Ka"	0.02	**0.90**	0.11	-0.03	-0.13
Case particle "To"	0.31	**0.71**	0.00	0.05	-0.02
Final particle "Ka"	-0.16	**0.61**	0.19	0.07	**0.41**
Connective particle "Shi"	0.00	**0.60**	0.21	0.08	-0.12
Incidental particle "Mo"	**0.42**	**0.56**	0.02	0.03	0.03
Connective particle "Keredo"	**0.47**	**0.56**	0.00	-0.05	-0.02
Auxiliary particle "Tari"	0.24	**0.53**	-0.29	0.00	-0.04
Final particle "Na"	-0.05	**0.51**	0.28	0.32	-0.04
Auxiliary verb "Rashii"	-0.06	0.34	0.30	-0.03	-0.12
Final particle "Yo"	0.10	-0.11	**0.77**	0.04	0.28
Final particle "Jan"	-0.08	0.00	**0.67**	0.05	-0.13
Auxiliary verb "Da"	0.32	0.31	**0.56**	-0.07	-0.07
Final particle "Sa"	-0.15	0.36	**0.52**	0.04	-0.24
Final particle "Mono"	0.14	-0.16	**0.51**	**0.46**	0.05
Auxiliary verb "Tai"	-0.21	**0.41**	**0.47**	0.09	0.03
Final particle "Ke"	0.04	0.08	**0.42**	-0.08	-0.10
Auxiliary particle "Shika"	0.13	0.08	0.39	0.00	0.12
Auxiliary particle "Kurai"	0.27	0.25	0.29	-0.09	0.07
Auxiliary verb "Zu"	-0.07	0.08	0.18	**0.88**	0.15
Auxiliary verb "Toru"	-0.10	-0.03	0.18	**0.83**	0.09
Auxiliary verb "Ya"	0.04	0.24	-0.33	**0.58**	-0.15
Auxiliary verb "Desu"	0.13	-0.17	-0.17	-0.01	**0.98**
Auxiliary verb "Masu"	0.11	-0.04	-0.21	0.07	**0.85**

Table 2: Results of factor analysis of frequently appearing function words in the NUCC

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Female 10s	-0.04	0.32	1.06	0.35	-0.14
Female 20s	-0.16	0.19	0.18	-0.09	-0.12
Female 30s	-0.03	0.36	-0.16	-0.19	-0.03
Female 40s	0.05	-0.05	-0.50	-0.05	0.17
Female 50s	-0.34	-0.83	-0.80	0.07	-0.31
Female 60s	0.97	-0.34	0.04	-0.29	0.46
Female over 70s	1.27	-0.59	0.01	0.20	-0.53
Male 10s	0.22	0.76	1.39	2.82	0.14
Male 20s	-0.13	-0.06	0.30	0.53	0.17
Male 30s	-1.56	-1.29	-1.33	-0.31	-0.52
Male 40s	-0.02	-0.17	-0.71	0.24	0.73
Male 50s	-0.64	-0.86	-0.54	-0.40	0.16
Male 60s	0.26	-0.64	-0.58	0.12	0.87

Table 3: Average factor scores for each gender / age category in the NUCC

4. Factor Analysis for Utterance Styles

To extract the typical utterance styles of Japanese daily conversation, a factor analysis for frequencies of particles and auxiliary verbs was performed. Because of statistical limitations, the top 50 most frequent function words (particles and auxiliary verbs) were selected and 50 dimensional word frequency vectors were extracted for each speaker in each scene. The rotation method used was Promax and a parallel analysis was performed to determine the number of factors. As a result, five factors were identified. The resultant factor scores are shown in Table 2; the bold font signifies cells whose factor scores exceeded 0.4.

In order to investigate the meanings of each factor, the average factor scores were calculated as Table 3 in each of the categories from Table 1. The factor score shows the relationships between each factor and each data set. If a factor score is regularly high in some data sets, this suggests the correlation of the factor and the data sets.

The five factors corresponded with the frequently appearing utterance patterns in Japanese daily conversation in the NUCC. The characteristics and naming of each factor are as follows:

Factor 1: This factor includes general function words such as the case particles “Ga,” “No,” “Wo,” “Kara,” and “Ni.” Therefore, Factor 1 reflects neutral general usage in Japanese utterances. However, Factor 1 also includes some feminine characteristic words such as the final particles “Wa” and “No” as well as informal words frequently used by young speakers such as the auxiliary verbs “Chau” and “Tuu.” This combination of “neutral,” “feminine,” and “youth” characteristics may be occurring because of the bias of the NUCC. Table 1 clarifies that the NUCC includes more feminine and youth usage of utterances characteristically. Therefore, Factor 1 may represent “Average style in NUCC.”

Table 3 shows that this factor has a strong relationship among older females. This result may suggest that the traditional feminine utterance style is only applicable for older females in real conversation in modern Japan.

Factor 2: This factor includes such auxiliary particles as “Ka” and “Tari,” as well as the case particle “To,” the connective particle “Shi,” and the incidental particle “Mo.” These particles have the common functions of juxtaposing and continuation. This may reflect an utterance style of continued speaking without specifying the end of the sentence. This factor also includes the final particles “Ka” and “Na.” These two particles show some nuances of the interrogative form. These utterance styles may relate to both avoiding assertions and speaking in an ambiguous way. Above “Ka” and “Tari,” also have been utilized in similar way. This may be a result of some pragmatic strategy employed to avoid collisions and to enhance empathy. Therefore, Factor 2 is referred to as “Avoid affirmation style.”

This factor is commonly related to young females (10s, 20s, and 30s) and also to10s males. It may be characteristic of young female utterance styles in modern Japan. However, the 10s-male category includes only two people in the NUCC and therefore it cannot be concluded that Factor 2 is related to the young male demographic.

Factor 3: This factor includes such final particles as “Yo”, “Jan,” “Sa,” “Mono,” and “Ke.” These may characteristically reflect informal, frank communication styles. Moreover, Table 3 shows that this factor strongly related to 10s females and males. Therefore, Factor 3 is referred to as “Frank teenager style.” Although factor 1 also includes frank style, the differences from factor 1 are gender free, youth only, and separation from general utterance style.

Factor 4: This factor includes the auxiliary verbs "Zu," "Toru," and "Ya." Although the auxiliary verb "Zu" is a somewhat general word, "Toru" and "Ya" are characteristically used in various dialects. In the NUCC, some speakers also have dialect tones, and those may be reflected on this factor. Therefore, it was labeled "Dialect style."

This factor is strongly related to 10s male in Table 3 because one of the two 10s male speakers has strong dialect tone. However it cannot be generalized because of too small sample size.

Factor 5: This factor includes the auxiliary verbs "Desu" and "Masu." These are clearly related to Japanese honorific utterance styles. Therefore, Factor 5 was referred to as "Polite style."

There is no certain tendency in the correlating categories of this factor in Table 3. Honorific utterance styles are dependent on the social relationships between the speaker and the listener. Therefore, the categories of gender and age seem not to be related meaningfully in this factor.

The results from the examination of the NUCC were compared to those observed in the utterance styles in Japanese fictional texts (Murai, 2017A, 2017B), and three of these factors also appeared in the fictional texts: "Frank style," "Kansai dialect style," and "Polite style." In the cases of those three factors, the included words are not exactly the same, but they are very similar between real and fictional utterances. The "Average style in NUCC" seems to be combination of "Neutral style" and "Feminine style" in the factors of fictional utterances.

Though in previous research seven fictional utterance styles were obtained, "Crude style" and "Hearsay style" have not been observed in the real-life daily conversation corpus. "Crude style" often reflects hostile relationships between fictional characters and therefore would not appear in the experimental daily conversation apart from rare situations of quarrel. "Hearsay style" is frequently used in fictional conversation in order to diversify narrative forms. However, such diversification may not be necessary in everyday conversation.

On the other hand, "Avoid affirmation style" has not been observed in the fictional utterance corpus. It may be a new utterance style in modern Japan. Therefore, fictional writers may not recognize this utterance pattern. Instead that, fictional writers adopt traditional feminine speech style for their fictional feminine characters. However, traditional feminine speech style is used mainly in over 60s (factor 1) in real corpus data. This result would be help to understand the time span of utterance style change.

5. Conclusion and Future Work

The characteristics of fundamental utterance styles in Japanese daily conversation were analyzed by a factor analysis method based on the NUCC. As a result, five factors ("Average style in NUCC," "Avoid affirmation style," "Frank teenager style," "Dialect style," and "Polite style") were extracted quantitatively.

Compared to fictional utterance styles, "Avoid affirmation style" is unique for real-life daily conversation. On the other hand, "Crude style" and "Hearsay style" did not appear in that corpus.

Although the similarities between the fictional corpus and the NUCC support the validity of the result, the factors were likely affected by the bias of the corpus. It would be desirable to utilize a speaker-balanced daily conversation corpus for more precise analysis.

Moreover, knowledge of the relationships between the speakers and the listeners would be useful for obtaining detailed characteristics of utterance styles.

6. Acknowledgements

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'.

7. Bibliographical References

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

Kurosawa, A. (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, pp. 3–11. (In Japanese)

Miyazaki, C., Hirano, T., Higashinaka, R., Makino, T., Matsuo, Y., & Sato, S. (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)

Murai, H. (2017A). Situational Effects on Functional Word Frequencies within Conversational Sentences in Japanese Novels. Proceedings of JADH Annual Conference 2017, pp. 40—42.

Murai, H. (2017B). Characteristics of Utterances in Japanese Fiction-writing. IJCAI 2017, the 2nd. International Workshop on Language Sense on Computer, pp. 6—10.

Seto, K. & Kishi, Y. (2015). Construction of a Dialogue System Using a Speech Type of Estimation by Adjacency. Proceedings of Information Processing Society of Japan 2015, pp. 131—132. (In Japanese)

Shen, R., Kikuchi, H., Ohta, K., & Mitamura, T. (2012). Towards the text-level characterization based on speech generation. Journal of Information Processing Society of Japan, 53(4), pp. 1269–1276. (In Japanese)

8. Language Resource References

Fujimura, I., Chiba, S., & Ohso, M. (2012). Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a Lexical Profiling Approach to Comparing Spoken and Written Corpora, Proceedings of the VIIth GSCP International Conference. Speech and Corpora, pp. 393—398.