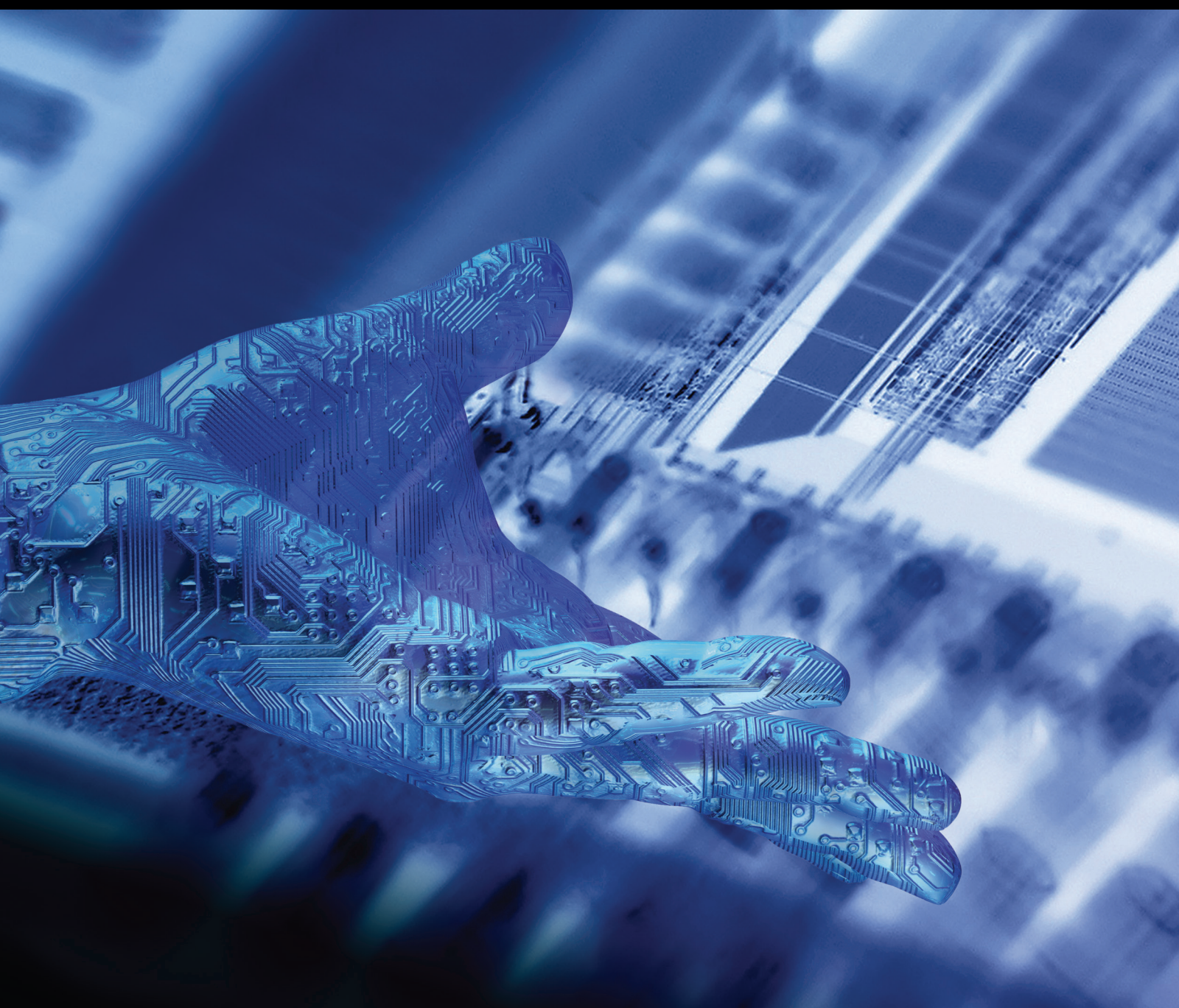
Advances in Human-Computer Interaction

Language Sense and Communication on Computer

Lead Guest Editor: Akinori Abe
Guest Editors: Rafal Rzepka and Michal Ptaszynski



Language Sense and Communication on Computer

Advances in Human-Computer Interaction

Language Sense and Communication on Computer

Lead Guest Editor: Akinori Abe
Guest Editors: Rafal Rzepka and Michal Ptaszynski

Editorial Board

Contents

Hindawi
Advances in Human-Computer Interaction
Volume 2019, Article ID 3081602, 2 pages
https://doi.org/10.1155/2019/3081602

Editorial

Language Sense and Communication on Computer

Akinori Abe,[1] **Rafal Rzepka**,[2] **and Michal Ptaszynski**[3]

[1]*Chiba University, Chiba, Japan*
[2]*Hokkaido university, Sapporo, Japan*
[3]*Kitami Institute of Technology, Kitami, Japan*

Correspondence should be addressed to Akinori Abe; ave@ultimavi.arc.net.my

Received 11 March 2019; Accepted 11 March 2019; Published 2 May 2019

We have roughly defined the "language sense" as an expression that underlines affective or psychological aspects of language. Today in the AI field, the ability (power) of mathematical calculation and data analysis will be superior to human power. However, for the emotional aspect (KANSEI in Japanese) of processing, still computers cannot deal with such problems properly. Accordingly in this special issue, we would have liked to deal with several problems related to language sense on computers and applications dealing with *language sense* in written texts and dialogues.

For this special issue we could accept eight papers.

The paper by M. Ogasahara et al. deals with rumors on Twitter. In fact, Twitter will be one of the good tools to express our ideology and philosophy as well as simple information. They analyzed rumor on Twitter during the 2011 Great East Japan Earthquake. They showed the difference in expression changes between diffused rumor tweets and nondiffused rumor tweets during the Great East Japan Earthquake in 2011. The contrast between the few expression changed in diffused tweets and the many such changes in nondiffused tweets implied the existence of information filters on Twitter, namely, hubs. Thus they showed language expression change as language sense.

The paper by T. Matsuki and J. Nakamura deals with the effects of the values and attitudes of retail and service industry employees on employee satisfaction (ES) and identified differences between regular and nonregular employees. They concluded employee values affected ES; the values of regular and nonregular employees are not significantly statistically different. They also showed that keywords of free answer comments implied the values of both features.

In the paper by T. Hayashi and Y. Ohsawa, they discuss what kinds of data should be acquired to understand situations of belief drift (BD). They showed that even though the terms used to explain events or problems differ, since the framework of thought and understanding are relatively the same, the Variable Labels necessary for understanding the state of BD attained higher commonality. Thus their results suggest that, even if the terms used to explain the state of BD differ, the data acquired to understand BD are common.

In the paper by T. Akimoto, his basic idea is that stories are representational elements forming an agent's mental world and are also living objects that have the power of self-organization. Accordingly, he developed this idea by discussing the complexities of the internal structure of a story and the organizational structure of a mental world. In particular, he classified the principles of the self-organization of a mental world into five types of generative actions, i.e., connective, hierarchical, contextual, gathering, and adaptive. An integrative cognition is explained with these generative actions in the form of a distributed multiagent system of stories.

In the paper by N. Komiya and J. Nakamura, based on the idea that a word can carry different meanings for different people, they conducted conjoint analysis. Particularly it was applied to assess preferences for various words describing integrated resorts (IR) including casinos, to be introduced in Japan in the future. They discussed how the participants understood particular words (e.g., a specific casino's place name or wording regarding restrictions on betting) that define the characteristics of a casino, as well as how casino-related words influenced participants' preferences.

The paper by M. Uther et al. investigated user experiences from 117 Finnish children aged between 8 and 12 years in a trial of an English language learning programme that used automatic speech recognition (ASR). They used measures that encompassed both affective reactions and questions tapping into the children's sense of pedagogical utility. They also tested children's perception of sound quality and compared reactions of game and nongame-based versions of the application. Their results showed that children expressed higher affective ratings for the game compared to nongame version of the application. Children also expressed a preference to play with a friend compared to playing alone or playing within a group. Children found assessment of their speech useful although they did not necessarily enjoy hearing their own voices. Their results can be discussed in terms of the implications for UI user interface (UI) design in speech learning applications for children.

H. Murai's paper analyzed the characteristics of utterances in Japanese novels. Based on the attribute (e.g., the speaker, listener, relationship between the speaker and listener, and gender of the speaker) 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. His 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 datasets in BCCWJ (Balanced Corpus of Contemporary Written Japanese) 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 utilizable in speaker identification tasks, automatic speech generation tasks, and scientific interpretation of stories and characters.

The paper by R. Jucks et al. reported a study that manipulates an SDS's (spoken dialogue system) word use with regard to politeness. In an experiment, 58 young adults evaluated the spoken messages of our self-developed SDS as it replied to typical questions posed by university freshmen. The answers were formulated either politely or rudely. Dependent measures were both holistic measures of how students perceived the SDS as well as detailed evaluations of each single answer. Their results showed that participants evaluated not only the content of rude answers as being less appropriate and less pleasant than the polite answers, but also the rude system as less accurate. Lack of politeness also impacted aspects of the perceived trustworthiness of the SDS. They concluded that users of SDS expect such systems to be polite, and they then discussed some practical implications for designing SDS.

Thus several features of language sense were analyzed. Reviewing papers in this special issue, a strategy or technique for a literary work generation can be suggested. In addition, hidden factors in spoken or Twitter language have been revealed. By using results, for instance, a very natural, sophisticated, and smart chat or Twitter system can be constructed.

Conflicts of Interest

The editors declare that they have no conflicts of interest regarding the publication of this special issue.

Akinori Abe
Rafal Rzepka
Michal Ptaszynski

Hindawi
Advances in Human-Computer Interaction
Volume 2019, Article ID 5103840, 8 pages
https://doi.org/10.1155/2019/5103840

Research Article

How Did Rumors on Twitter Diffuse and Change in Expression? An Investigation of the Process of Spreading Rumors on Twitter during the 2011 Great East Japan Earthquake

Morihiro Ogasahara,[1] **Hirotaka Kawashima**,[2] **and Hiroyuki Fujishiro**[3]

[1]*Faculty of Sociology, Kansai University, Osaka, Japan*
[2]*National Institute of Science and Technology Policy, Tokyo, Japan*
[3]*Faculty of Social Sciences, Hosei University, Tokyo, Japan*

Correspondence should be addressed to Morihiro Ogasahara; m36ogasahara@gmail.com

Received 29 June 2018; Accepted 16 October 2018; Published 23 January 2019

Guest Editor: Akinori Abe

Twitter has been emerging as a major communication channels during disasters. The characteristics of Twitter—quickly and widely diffusing information and allowing every user to be an information transmitter—can be effective for problem solving in anxious and ambiguous situations such as disasters; however, false rumors on Twitter can be a serious problem. Rumor research has suggested that rumors are a kind of collective sense-making when mass media cannot provide people with enough information (e.g., during disasters). Furthermore, the expression of the rumor changes during the process of spreading. This study investigated the data of 187 thousand tweets related to the Cosmo Oil rumor during the Great East Japan Earthquake of March 2011 and analyzed the change in Twitter expressions and the collective sense-making process during this catastrophe. The results of this study suggest that collective sense-making is rare in diffused tweets, partly because of the gatekeeping role of hubs (i.e., users who have many followers on Twitter). Rumor discussion on Twitter during disasters might be suitable for broadcasting static information than for collective sense-making.

1. Introduction

Twitter is a microblogging service where users post messages up to 140 characters. In 2018, it had 326 million monthly active users worldwide. Because there are no limits to the number of followers (i.e., readers of specific Twitter users' feeds), popular Twitter users can distribute their Twitter posts ("tweets") to a large number of people. For example, President Trump of the United States had, as of 2018, over 55 million followers. Another feature of Twitter communication is the ease with which posts can be forwarded to other users ("retweets"). Twitter users can allow their followers to read other user's tweets using only two clicks on the "retweet" button for each tweet that they read. Because of these characteristics, specific tweets can spread to millions of people over a very short period on Twitter [1].

Twitter is emerging as one of the main communication channels during disasters. Using Twitter, authorities were able to directly distribute necessary information to people, including disaster victims, and people could actively exchange disaster-related information during the flooding of Red River Valley in 2009 [2], the Great East Japan Earthquake in 2011 [3], and Hurricane Sandy in 2012 [4]. However, Twitter also spreads misinformation and false rumors. During the 2011 Great East Japan Earthquake, a prominent false rumor claiming that rainfall contained harmful material began to spread right after the explosion of the LGP tanks at Cosmo Oil Co., Ltd. (known as the "Cosmo Oil rumor"). Furthermore, during Hurricane Sandy in 2012, fake images of the hurricane widely and quickly diffused worldwide on Twitter. False rumors can confuse people and authorities, as well as interfere with managing the disaster. Governments

have begun struggling with false rumors spreading via social media, including Twitter.

Rumor research suggests that rumors are a form of collective sense-making process rather than mere transmission of information (or misinformation) [5–7]. During disasters, people experience anxiety and their need for information explodes; however, mass media sometimes cannot supply enough information to meet the suddenly huge demand. In the absence of formal information about a highly anxiety-inducing and ambiguous situation, people become motivated to share and evaluate information in order to explain the situation. Rumors have been defined as "unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger or potential threat, and that function to help people make sense and manage risk" [8].

From this point of view, rumor should not necessarily be excluded during a disaster. In order to foster the potentially positive uses of rumor and inhibit the negative ones, understanding the dynamics of rumor discussion on Twitter would be exceedingly important. This study investigates the pattern of expression changes over time and the collective sense-making process on Twitter using real tweet data about the Cosmo Oil rumor.

2. Related Work

2.1. Expression Changes in Rumor. Rumor researchers have shown that the expression of rumors changes over time and have studied the patterns and mechanism of such change. Allport and Postman conducted laboratory studies on expression changes in rumors. They instructed participants to describe a drawing and transmit that description through a chain of participants without discussion [9]. Rosnow and Fine also conducted field studies using false rumors, such as the "Paul McCartney is dead" rumor. These studies identified four patterns of rumor expression change: *leveling*: "the loss of detail and the reduction in length at each successive transmission"; *adding*: the "addition to rumor content in the form of new material or additional details"; *sharpening*: "the accenting and highlighting of certain details in the rumor message"; and *assimilation*: "the shaping of rumor content—through leveling, adding, and sharpening—so as to be in greater accord with personal cognitive schemas" [8].

RQ1: What kind of expression change is observed in Cosmo Oil rumor on Twitter over time?

2.2. Rumor as Collective Sense-Making. Laboratory studies of rumor such as Allport and Postman's were criticized because the listener was not allowed to seek clarification or cross-examine the speaker, whereas such interactions do tend to be observed in real-life rumor situations [5]. Shibutani proposed the perspective of rumor as a form of collective sense-making: when formal information is absent and people are in an anxiety-inducing and ambiguous situation (e.g., during a disaster), people tend to compensate by informally interpreting the situation [7]. Although the perspective of rumor as collective sense-making has been supported through many field studies, it has, until now, lacked quantitative evidence because of the difficulties in collecting data on real-life rumor transmission.

Computer-mediated communication (CMC) has enabled researchers to record all interactions during rumor transmission, thus providing the necessary quantitative evidence for rumor as a collective sense-making process. Bordia and Rosnow [6] devised a general scheme for coding and analyzing statements in rumor transmission on a CMC network, and Bordia and DiFonzo [5] modified this scheme to create the Rumor Interaction Analysis Systems (RIAS). They identified 14 types of rumor statements through an analysis of 14 rumor discussions in CMC discussion groups using RIAS and found that the most frequent statement in the rumor discussions was the sense-making type (29.4%). Additionally, they divided the progression of rumor discussions over time into quarters by dividing the number of postings by 4 and derived the main types of statements in each quarter. They found that interrogatory statements peaked during the first quarter, while sense-making statements peaked during the third quarter. Their findings suggest that rumor discussion is a collective sense-making process and that the types of statements made in such discussions change over time. On the other hand, Oh et al. [10] analyzed tweets related to the Haichi Earthquake in 2010 using the RIAS and found that sense-making tweets were very rare. They posited that the Twitter interface, which limits tweets to 140 characters, might make such sense-making statements difficult.

RQ2: What statement types are observed in Cosmo Oil rumor spread over Twitter and how do the amount of these statement types change over time?

2.3. Rumor Diffusion on Twitter. Since Twitter has a distinct user-follower network and a distinct message forwarding function (retweets), rumor transmission on Twitter might have different features from face-to-face interaction or CMC discussion groups. Mendoza et al. [11] observed there are far more denials and questions about tweets of false rumors than about tweets related to confirmed truths and suggested that false rumors can be detected using an aggregate analysis of tweets. However, the number of denying/questioning tweets does not necessarily equate to the number of readers.

Kwak et al. [1] noted that retweets generally reach an average of 1,000 users, regardless of the number of followers of the user who made the original tweet. Arif et al. [12] examined rumors on Twitter during a hostage crisis in Sydney in 2015 and proposed that tweets from accounts with low followers could still diffuse widely when they have a high number of derivative tweets (i.e., identifiable downstream tweets that are copies of the original tweets, including retweets and small amount of rewords). Hence, the number of retweets might be deemed a rough estimate of the number of readers of the tweets. Furthermore, it would be difficult for rumor tweets with few retweets (which would imply few readers) to contribute to the collective sense-making on Twitter. Thus, rumor tweets with many retweets (*diffused* rumor tweets) and ones with few retweets (*nondiffused* rumor tweets) should be examined separately.

Table 1: Statement coding category and its definition.

Category	Definition
Authenticating (Au)	Statements expressing the speaker's attempt to add credibility to what he or she was saying.
Sense-making (Sm)	Statements reflecting attempts to solve the problem of whether or not the rumor is true, including providing information to solve the problem.
Emotional (Em)	Emotionally charged expressions that include both positive and negative feelings.
Interrogatory (I)	Questions seeking information.
Directive (Dr)	Statements that suggests a course of action.
Unrelated (Ur)	Statements that are not relevant to the original rumor.
Uncodable (Uc)	Statements that are not able to be categorized.

RQ3: What is the difference between *diffused* rumor tweets and *nondiffused* ones in terms of expression changes and statement types in Cosmo Oil rumor?

3. Methods

3.1. Background of the Cosmo Oil Rumor. The Great East Japan Earthquake hit eastern Japan at 14:46 (JST) on March 11, 2011. The magnitude 9.0 earthquake caused an explosion at the LPG tanks of Cosmo Oil Co., Ltd. in Ichihara city, Chiba prefecture at 17:04 and various TV and radio stations reported on the explosion repeatedly. After the explosion, a tweet that said "rain with harmful materials is falling" began to rapidly diffuse on Twitter, chain mail, and mixi (a Japanese domestic social networking service). On March 12, Cosmo Oil Co., Ltd., the authorities, and mass media officially denied the Cosmo Oil rumor, causing it to subside almost instantly.

3.2. Data. Twitter Japan gave us permission to use all tweet data posted from March 11 to March 17, 2011, in the workshop for big data on the Great East Japan Earthquake—Project 311 [13]. It contains the following information on each tweet: tweet ID, account ID, timestamp, and tweet text. Any tweets deleted before the workshop began were not included in the data. We extracted the tweets related to the Cosmo Oil rumor from the all tweet data using the keyword "cosmo." A dataset of 187,202 extracted tweets (*Cosmo Oil dataset*) was used in this study.

Authors and three research collaborators read all the tweet contents in the *Cosmo Oil dataset* and extracted notable tweets for understanding the dynamics of the change in expression and diffusion process of Cosmo Oil rumors. We operationally defined *diffused* tweets as tweets that had been retweeted over 100 times; 87 tweets were classified as *diffused,* with all other tweets being classified as *nondiffused*. The total number of tweets and retweets of these *diffused* tweets was 75,831. *Diffused* tweets and their retweets account for 40.5% of the *Cosmo Oil dataset*.

3.3. Coding. Coding was performed with reference to the classification by Arif et al. [12] and the RIAS [5], who employed five mutually exclusive categories—affirm, deny, neutral, unrelated, and uncodable—to capture the overall trend of the rumor discussion on Twitter. Bordia and DiFonzo [5] employed 14 categories of statements to RIAS to analyze the rumor discussion in CMC discussion groups. They were prudent, apprehensive, authenticating, interrogatory, providing information, belief, disbelief, sense-making, directive, sarcastic, wish, personal involvement, digressive, and uncodable. Oh et al. [10] revised RIAS to analyze the rumor discussion on Twitter and dropped the seven categories of digressive, personal involvement, wish, sarcastic, apprehensive, providing information, and sense-making, because they were very rare in their dataset of tweets related to the Haiti Earthquake. They also replaced apprehensive statements with emotional ones, which included both positive and negative dimensions.

First, we coded the attitude of the rumor tweets in the *Cosmo Oil dataset* with the same categories used by Arif.et al.—affirm, deny, neutral, unrelated, and uncodable—to capture the overall stance of the Cosmo Oil rumor discussion. A tweet coded as "affirm" implies that it endorses or affirms the rumor whereas a tweet coded as "deny" disputes or refutes the rumor. A tweet coded as "neutral" neither directly affirms nor denies the rumor but is still related to the story.

Second, we coded the statements of the tweets to capture the sense-making process in the Cosmo Oil rumor discussion in detail based on the following seven categories: authenticating, sense-making, emotional, interrogatory, directive, unrelated, and uncodable, as shown in Table 1. These categories were referring to RIAS and were not necessarily mutually exclusive. The statements that may have been coded as belief, disbelief, and prudent, using RIAS were coded as affirm, deny, and neutral, respectively in the first step of our coding procedure as noted above. As Oh et al. analyzed, we replaced apprehensive statements in RIAS by emotional and omitted the three categories of sarcastic, wish, and personal involvement from RIAS because they were very rare in the *Cosmo Oil dataset*. The providing information category was also omitted. Because information providing statements were presumed to be responses to interrogatory statements [5], it was difficult to distinguish between responses to interrogatory statements and mere information transmission in the rumor discussion on Twitter. We retained the sense-making category in RIAS because our research interest focused on the sense-making process in the rumor discussion.

In this study, we recruited two coders from among our university students, instructed them how to code tweets using the coding manual, and let them code all the *diffused* tweets and the notable *nondiffused* tweets introduced in this

paper later. The intercoder reliability: Cohen's κ [14] for classification of the attitude of the rumor tweets was 0.73 and that for statements of the rumor tweets was 0.84. These reliabilities could be judged as substantial agreement and almost perfect agreement, respectively [15].

3.4. Dividing the Diffused Tweets. To capture expression changes over the course of the Cosmo Oil rumor discussion, we divided *diffused* tweets into "quarters," with reference to the analysis method proposed by DiFonzo and Bordia [8]. We divided all *diffused* tweets except unrelated (Ur) ones into quarters by dividing the number of *diffused* tweets chronologically by 4, and then counting the number of each type of attitude and statement in each quarter. While DiFonzo and Bordia divided and analyzed all posts (7 to 54 posts) in the 14 rumor discussions on computer discussion groups, it was quite difficult to divide and count each type of attitude and the statements of all tweets (187,202 tweets) in the *Cosmo Oil dataset* in the same way. However, because the percentage of diffused tweets and their retweets was not small in the *Cosmo Oil dataset* as noted before, and because tweets with a large number of retweets would have been read by far more people than would tweets with fewer retweets [1], we judged that dividing only *diffused* tweets would be acceptable in this research.

4. Results

In this chapter, we introduce the texts of tweets that featured communication about the Cosmo Oil rumor. The texts includes the timestamp, tweet text, the number of retweeted (number plus "RT" in parentheses), and the statement category. User information is referred to only when it is necessary.

4.1. Diffused Rumor Tweets. The first *diffused* tweet of the Cosmo Oil rumor occurred at 6 pm (Tweet A). We identified it as the *original rumor tweet* and named the main text of the tweet as the *core rumor text* (underlined part of the original rumor tweet).

Tweet A (directive)

@username (03-11 18:43:40)

[Please share] To residents around Chiba city! Due to the Cosmo Oil explosion, harmful material has become adherent to clouds, and is falling with rain. When you go outside, bring an umbrella or a raincoat.Don't let your body be exposed to rain!!! (1,759RT)

After the original rumor tweet was posted, the number of tweets related to the Cosmo Oil rumor exploded. A total of 18 *diffused* rumor tweets affirmed the rumor. Since two of these tweets referred to the rumor as related information for other discussions, we concluded that 16 *diffused* tweets contributed to the discussion of the Cosmo Oil rumor. Interestingly, all 16 *diffused* rumor tweets quoted the full text or core rumor text of the original rumor tweet without changes. Since the downstream tweets kept the text of the original rumor tweet almost intact, it seems evident that leveling, sharpening, and assimilation did not occur in the discourse of the Cosmo Oil rumor. Adding was observed in authenticating statements in some tweets. Another feature of the *diffused* affirming tweets was that these 16 tweets contained only two statement categories: authenticating and directive.

Table 2: Numbers of attitudes in each quarter.

	Affirm	Deny	Neutral
Q1 (n=20)	17	3	0
Q2 (n=19)	0	19	0
Q3 (n=19)	0	19	0
Q4 (n=19)	1	17	1

Tweet B (03-12 00:12:30, directive)

Share more! (core rumor text), Please let everyone know this!! (1,342RT)

Tweet C (03-11 19:58:50, authenticating and directive)

Fuji television said so. RT (original rumor text). (614RT)

Tweet D (03-12 09:23:42, authenticating and directive)

The information from the factory worker. Take care of outing and don't reveal your skin. (core rumor text) Please let everyone know this!! (2,055RT)

It took about 16 hours from the posting of the original rumor tweet for the first *diffused* rumor tweet denying the rumor to appear ("NHK" in tweet E is Japanese national public broadcasting organization).

Tweet E (03-12 11:00:14, authenticating and sense-making)

[Can I say a few words?] It is a HOAX that exposure to rain is dangerous, harmful material is mixed in rain due to the explosion. NHK also said so. Don't share the rumor, it bothers the Cosmo staff. I heard that when the oil burns, it will produce only carbon dioxide. (2,055RT)

When Cosmo Oil Co. Ltd., authorities, and Asahi Shimbun (a daily newspaper in Japan) officially denied the Cosmo Oil rumor, denial tweets spread widely in an instant. The most retweeted deny tweet was posted by Urayasu City Hall in Chiba prefecture.

Tweet F (03-12 15:31:53, authenticating and directive)

@urayasu_koho

A false chain mail has been sent, saying that harmful material is falling with the rain due to the LPG tank explosion at the Cosmo Oil Chiba refinery. We have asked the Fire, Earthquake and Disaster Defense Division of Chiba prefecture and confirmed it to be false. Please act only with accurate information. (21,078RT)

We arranged 77 *diffused* rumor tweets in chronological order and divided them into quarters; then we counted each type of attitude (Table 2) and statement (Table 3). Unrelated (Ur) *diffused* tweets (10 tweets) were excluded from this analysis. Because there were no uncodable tweets regarding attitude, and because the only uncodable tweet about statement was a tweet that simply expressed denial of the rumor, the property of which was already considered in Table 2, the "uncodable" column was omitted from both tables.

Table 3: Numbers of statements in each quarter.

	Au	Sm	Em	I	Dr
Q1 (n=20)	7	1	0	0	20
Q2 (n=19)	17	1	0	0	8
Q3 (n=19)	17	2	0	0	7
Q4 (n=19)	15	0	0	1	13

Table 2 shows that (1) almost all affirmative tweets are concentrated in Q1, (2) denial tweets dominated from Q2 to Q4, and (3) there are very few neutral tweets in any of the quarters. These patterns directly reflect the existence of an official denial. There were no official denials to the Cosmo Oil rumor in Q1, whereas Cosmo Oil Co., Ltd. denied the rumor in Q2 and Urayasu City Hall and Asahi Shimbun denied it in Q3.

Table 3 also suggests a similar pattern. Less than half of the statements were coded as authenticating (Au) in Q1, while most of the statements in Q2 to Q4 were classified as such. This change is consistent with the finding of Oh et al. [16], who noted that information with no clear source was the most important rumor-causing factor on Twitter. Second, there were very few sense-making and interrogatory statements and no emotional statements in any of the periods. The ratio of directive statements in all tweets is the highest in Q1, which then dropped in Q2 to Q3 before increasing again in Q4.

4.2. Nondiffused Rumor Tweets. In this section, we analyze the changes of expression, attitude, and statements of *nondiffused* rumor tweets using the same framework as the *diffused* ones. However, it is quite difficult to count the number of each type of change over 100 thousand tweets. Thus, we introduce typical tweets in the *Cosmo Oil dataset* for each type of change.

While adding was the only type of expression change observed in *diffused* rumor tweets, we observed other types in the *nondiffused* rumor tweets.

Tweet G (03-11 18:00:26, directive)

I heard that water solution is flying apart as a consequence of the Cosmo Oil fire in Ichihara city, Chiba prefecture. Absolutely do not get exposed to rain, prevent it with a mask. I also heard that Shinkansen of the Tohoku, Joestu, and Akita lines will be stopped all day today.(7RT)

The timestamp of Tweet G is 43 minutes earlier than Tweet A: the original rumor tweet. Since Tweet G was posted earlier than Tweet A, and the content is quite similar, we presumed that Tweet G was the seed of the original rumor (*rumor seed tweet*). There are notable expression changes between the rumor seed tweet and the original rumor tweet: (1) the content concerning the "Shinkansen" (Japanese superexpress) in the seed rumor was omitted in the original rumor (leveling); (2) the expression "water solution" was changed into the more intense "harmful material" (sharpening); (3) references to clouds, umbrellas, and raincoats were added (adding); and (4) the text of the original rumor was more focused on the central theme of the rumor—that rainfall contained harmful material—than the seed rumor (assimilation). In short, all four types of expression change in rumor occurred from the seed rumor to the original rumor.

Tweet H (03-11 19:02:19, directive)

[Please share] Cosmo Oil in Itsui, Chiba prefecture exploded. Rain with harmful material (danger to the body) is falling due to the fire fighting. The air is also harmful. Kawasaki iron mill, too. Please don't go out except when necessary. It is dangerous to be within a 20 km radius of the explosion. (6RT)

Expression changes did not only occur from the seed rumor to the original rumor. Tweet H was posted after the original rumor tweet. References to umbrellas and raincoats were omitted from Tweet H (leveling), while references to a different danger zone, the potential harm caused by the air, and a description of the breadth of the danger zone were added (adding). The text of Tweet H is unified, as with the original rumor; however, Tweet H took a slightly different focus due to its differing context (sharpening).

Tweet I (03-11 21:13:36, authenticating)

[According to Chiba TV] I heard that the Cosmo Oil fire in Ichikawa city is safe because there are no leaks of toxic gas or harmful material. I saw the tweet "rain with harmful material is falling", but there are no such reports. (11RT)

As Tweet E shows, the first *diffused* deny tweet was posted 16 hours after the original rumor tweet. However, the first *nondiffused* deny tweet was posted about 1 hour later. Tweet I, a deny tweet with an authenticating statement, was posted about 2 hours later.

Tweet J (03-11 21:07:31, directive)

Although I do not know whether it is true or false, I heard that rain with harmful material is spreading. Prevent your skin from being exposed to rain with a raincoat or an umbrella. (0RT)

While one of the features of the *diffused* rumor tweets was that there were very few neutral tweets (Table 2), there were several neutral tweets among the *nondiffused* rumor tweets. These tweets often withheld judgment about the truth of the rumor and attempted to maintain neutrality just in case, even when they experienced doubt (Tweet J).

Tweet K (03-11 23:30:01, sense-making)

I don't want this tweet to be shared. If purified oil or LPG is burning only CO2 and water will be made. Even in the case of incomplete combustion, no harmful material will be made. RT [Forwarding] (original rumor text). (0RT)

Tweet L (03-11 20:57:59, interrogatory)

What is the information source for the tweet: "rain with harmful material is falling"? (0RT)

Tweet M (03-11 20:11:32, emotional)

I'm scared ... (original rumor text). (1RT)

In terms of statement type, we observed very few sense-making and interrogatory statements and no emotional statements among the *diffused* rumor tweets (Table 3). However, all the three types of statements were observed among the *nondiffused* rumor tweets. In Tweet K, the user rebutted the

rumor using his/her knowledge of chemistry. Many users also questioned the source of the information of the original rumor tweet (Tweet L). In Tweet M, the user expressed his/her fear on the rumor text straightforwardly.

5. Discussion

We assumed that rumor is a kind of collective sense-making and investigated tweet data about the Cosmo Oil rumor during the Great East Japan Earthquake in 2011 in order to clarify the pattern of expression changes over time and sense-making process of rumor on Twitter. In this section, we summarize our findings based on the three research questions.

Arguably, the most notable finding of this study is that both the pattern of expression change and the sense-making process of rumor differed substantially between *diffused* and *nondiffused* tweets (RQ3; Table 4). For *diffused* tweets, almost the entirety of the original rumor text or core rumor text was quoted in downstream tweets. This resulted in little expression changes, except for adding to the original text. On the other hand, all four types of expression change were observed in *nondiffused* tweets: leveling, adding, sharpening, and assimilation (RQ1).

Similarly, there seems to be no collective sense-making process in *diffused* tweets; attitudes to the rumor were only affirm or deny, and very few were neutral. Furthermore, very few sense-making statements and interrogatory statements, and no emotional statements were made. This pattern can be interpreted as a generally positive attitude towards rumor at the early stage. However, the pattern changes entirely once an official information source transmits accurate information. This pattern resembles the often passive audiences of mass media. The patterns in the *nondiffused* tweets are the opposite: various types of statements were observed, including sense-making, interrogatory, and emotional (RQ2). Users raised various questions and offered their thoughts on solving the problem, regardless of whether they believed the rumor to be true or false. This type of interaction exactly corresponds to a collective sense-making process. Users also easily expressed their rumor-related feelings of anxiety.

The question is, why is there such an enormous difference in patterns on Twitter? One of the reasons why certain tweets are *diffused* and *nondiffused* is follower distribution on Twitter. Since follower distribution on Twitter is largely skewed [1], hubs (users who have a much larger number of followers than ordinary users; e.g., president Trump or celebs) can influence the diffusion process of tweets. According to the survey report by Impress Corporation [17], Japanese Twitter users had an average of 54.4 followers in 2011, and 0.9% of Japanese Twitter users had more than 1,000 followers. Yasuda [18] analyzed the Cosmo Oil rumor tweets data and suggested that retweets by minihubs (twitter accounts with 1,000 or more followers) or hubs (twitter accounts with 10,000 or more followers) were critically important for the spread of rumor tweets, especially at the early stage of their diffusion. Further, she inferred that the hubs did not necessarily have the ability to discriminate truth from false rumor, as the hubs (e.g. fashion models or artists) were not specialists or authorities. In this research, Tweet F was posted by the Urayasu City Hall account, which had 14,408 followers in 2011. However, while the account that first posted Tweet E had a small number of followers (389 followers in 2018), one of the accounts that retweeted Tweet E was a hub with a significantly greater number of followers (16, 973 followers in 2018). Therefore, in the case of Tweet E, the hub seemed to enhance the diffusion of the tweet.

Since information sharing on Twitter tends to be motivated by reputation and efficacy [19], it seems plausible that hubs—whose tweets will be read by many strangers—would be more conscious of the content of their retweets than would ordinary users. Questioning (interrogatory), considering (sense-making), and being open to both pros and cons (neutral) are important factors in the collective sense-making process; however, these types of tweets seem to be avoided by hubs' followers, perhaps because of their high information processing load, especially during disasters. Hence, hubs on Twitter would be more likely to retweet a mere fact or conclusion than part of a complex and ambiguous sense-making process. If this interpretation is correct, the rarity of sense-making statements on Twitter is not only because of the 140-character limit for tweets, as suggested by Oh et al. [10], but also because of the role of hubs similar to opinion leaders in the two-step flow of communication [20, 21] or gatekeepers [22].

Twitter, at least in the information spaces composed by *diffused* tweets, does not seem suitable for dynamic collective sense-making. Rather, it seems more suited to exchanging static information during crises. The reason for the few expression changes in *diffused* tweets could be explained by this idea as well. Once a tweet attains the position of a *diffused* tweet, it becomes a kind of standard. The flood of modified tweets of these standards would confuse ordinary users, therefore hubs would be motivated to avoid to retweet them. Within such an information space, even if an ordinary user offers a significant question, useful insights, or expert knowledge, their information would be filtered by hubs and would not reach a broader audience. Instead, they would be buried in the flood of disaster-related information.

6. Conclusion

This study sheds light on the difference in expression changes between *diffused* rumor tweets and *nondiffused* rumor tweets during the Great East Japan Earthquake in 2011. The contrast between the few expression changes in *diffused* tweets and the many such changes in *nondiffused* tweets implies the existence of information filters on Twitter, namely, hubs.

However, there are still some limitations in this study. First, this study analyzed the data of only one rumor case on Twitter during a disaster. These findings should be further validated and replicated for other cases. Second, the classification criteria for *diffused* and *nondiffused* should be checked. The operational criteria in this study—over or less than 100 retweets—might be inappropriate. Third, the analysis of the *nondiffused* tweets should involve more than merely selecting example tweets. Indeed, we should analyze the *nondiffused* tweet data as a whole. Finally, because we did not have user

Table 4: Difference between *diffused* and *nondiffused* tweets.

	Expression change	Attitude	Statement type
Diffused	Small	Affirm/Deny/Very few neutral	Very few sense-making/interrogatory/emotional statements
Non-diffused	Large	Affirm/Deny/Several neutral	Several sense-making/interrogatory/emotional statements

network data on Twitter nor hubs' psychological motivation for retweets, we could not examine the gatekeeping role by hubs.

The more our daily communication depends on social media platforms including Twitter, Facebook, Instagram, etc., the more their influence will increase. The perspective that hubs influence not only which tweet is diffused but also which type of discourse emerges as a result can provide further insight into the dynamics of rumor discussion on Twitter. Further, our research perspective may be applicable not only to rumors but also to other types of communication, both on Twitter and other social media platforms.

Data Availability

The tweet data used to support the findings of this study have not been made available because the authors agreed about the terms and conditions stipulated by Twitter Japan, which have restricted the data set to be redistributed.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Twitter Japan for giving them permission to analyze the Japanese tweet data set from March 11 to March 17, 2011. They also thank the following research collaborators, who served as coders and discussion members during the course of this study: Yuzo Akakura, Asayo Iiduka, and Hiroshi Yamaguchi. This paper was partially supported by JSPS KAKENHI Grant no. 17K04179.

References

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 591–600, Rareigh, Calif, USA, April 2010.

[2] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on The Red: What hazards threat reveals about the social life of microblogged information," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '10)*, pp. 241–250, Savannah, GA, USA, February 2010.

[3] MIC, "White Paper 2011: Information and Communications in Japan" (Japanese), http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h23/html/nc143c00.html, 2011, (accessed October 3, 2018).

[4] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, pp. 729–736, Rio de Janeiro, Brazil, May 2013.

[5] P. Bordia and N. Difonzo, "Problem solving in social interactions on the internet: Rumor as social cognition," *Social Psychology Quarterly*, vol. 67, no. 1, pp. 33–49, 2004.

[6] P. Bordia and R. L. Rosnow, "Rumor rest stops on the information highway: Transmission patterns in a computer-mediated rumor chain," *Human Communication Research*, vol. 25, no. 2, pp. 163–179, 1998.

[7] T. Shibutani, *Improvised News: A Sociological Study of Rumor*, Bobbs-Merrill, Indianapolis, IN, USA, 1966.

[8] N. DiFonzo and P. Bordia, *Rumor Psychology: Social and Organizational Approaches*, American Psychological Association, Washington, DC, USA, 2007.

[9] G. W. Allport and L. Postman, *The psychology of rumor*, Henry Holt, Rinehart & Winston, New York, NY, USA, 1947.

[10] O. Oh, K. H. Kwon, and H. R. Rao, "An exploration of social media in extreme events: Rumor theory and twitter during the HAITI earthquake 2010," in *Proceedings of the 31st International Conference on Information Systems (ICIS '10)*, St. Louis, MO, USA, December 2010.

[11] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we RT?" in *Proceedings of the 1st Workshop on Social Media Analytics (SOMA '10)*, pp. 71–79, Washington, DC, USA, July 2010.

[12] A. Arif, K. Shanahan, F.-J. Chou, Y. Dosouto, K. Starbird, and E. S. Spiro, "How information snowballs: Exploring the role of exposure in online rumor propagation," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '16)*, pp. 466–477, San Francisco, Calif, USA, March 2016.

[13] Project, "The workshop for the big data of the Great East Japan Earthquake – Project 311," https://sites.google.com/site/prj311/, 2012, (accessed October 3, 2018).

[14] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[15] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[16] O. Oh, M. Agrawal, and H. R. Rao, "Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises," *MIS Quarterly: Management Information Systems*, vol. 37, no. 2, pp. 407–426, 2013.

[17] Impress et al., "The Report of Social Media Survey 2011" (Japanese), Impress, August 2011.

[18] Y. Yasuda, "Information dissemination in social media: hubs and demagogues," *Bulletin of the Faculty of Sociology, Kansai University*, vol. 45, no. 1, pp. 33–46, 2013 (Japanese).

[19] S. Y. Syn and S. Oh, "Why do social network site users share information on Facebook and Twitter?" *Journal of Information Science*, vol. 41, no. 5, pp. 553–569, 2015.

[20] E. Katz and P. F. Lazarsfeld, *Personal Influence*, The Free Press, NY, USA, 1995.

[21] E. Katz, "The two-step flow of communication: an up-to-date report on an hypothesis," *The Public Opinion Quarterly*, vol. 21, no. 1, pp. 61–78, 1957.

[22] P. J. Shoemaker and T. P. Vos, *Gatekeeping Theory*, Routledge, New York, NY, USA, 2009.

Hindawi
Advances in Human-Computer Interaction
Volume 2019, Article ID 8632892, 10 pages
https://doi.org/10.1155/2019/8632892

Research Article

Exploring the Types of Casinos Preferred in Japan via Conjoint Analysis of Relevant Words

Nozomi Komiya and Jun Nakamura

Shibaura Institute of Technology, 108-8548, 3-9-14 Shibaura, Minato-ku, Tokyo, Japan

Correspondence should be addressed to Nozomi Komiya; pa17005@shibaura-it.ac.jp

Received 24 June 2018; Revised 24 October 2018; Accepted 26 November 2018; Published 1 January 2019

Guest Editor: Akinori Abe

A word can carry different meanings for different people. Conjoint analysis was applied to assess preferences for various words describing integrated resorts (IR) including casinos, to be introduced in Japan in the future. We discuss how the participants understood particular words (e.g., a specific casino's place name or wording regarding restrictions on betting) that define the characteristics of a casino, as well as how casino-related words influenced participants' preferences. Implications for enhancing public understanding of casinos are explored in the conclusion.

1. Introduction

People understand words in a variety of ways, and the same words can be interpreted differently depending on one's knowledge, experience, and sensibility [1]. Errors and problems arise even in everyday circumstances due to differing interpretations of words. On the other hand, understanding can be improved by linking appropriate images with words. For example, consider an integrated resort (IR) that includes legal casinos. IRs contribute to the tourism industry in more than 141 regions worldwide and have produced a large economic effect [2]. For example, Macau has operated casinos as part of its tourism industry since 1847. Macao has turned casinos into entertainment facilities (with accommodation), revitalizing the casino market by transforming casinos to be more like resorts. In 2006, Macao became the most lucrative casino city worldwide, followed by Las Vegas [3].

In Japan, the Act on the Promotional Development of Areas for Specified Integrated Resort Facilities (the IR bill) was passed in 2016. During 2018, the National Assembly is expected to establish an IR implementation law outlining more specific regulations for the introduction of IRs. However, many people opposed to IRs [4]. Despite this resistance, it is noteworthy that the size of the gambling market in Japan (including public gambling and PACHINKO) exceeds 20 trillion yen per year. Therefore, Japan is akin to what may be called a "gambling center" [5].

Why has Japan failed to support casinos thus far, given that many people are familiar with the activities of casinos, such as large-scale public gambling and games of PACHINKO? In view of the abovementioned issues, the authors consider how words related to casinos have influenced people's preferences, i.e., how words may create negative or positive impressions of casinos.

The following chapter provides a review of the literature. The experimental details are discussed in the third chapter. The fourth section presents the results, and the discussion is set out in the fifth chapter. Finally, conclusions are drawn in the final chapter.

2. Literature Review

As part of an effort to rectify the negative image of public works policy in Japan, a follow-up survey of participants' impressions of various words related to public works was conducted [6]. Public works projects were originally conceived as contributing to the nation's relief efforts, safety, and disaster management. However, negative events such as accidents and delays in construction have taken a toll on people's impressions of public works. Tanaka et al. categorized

words used in the context of public works into four categories: (1) words directly describing public works; (2) Words directly related to public works; (3) words influencing public works; and (4) words indirectly related to public works. Tanaka et al. targeted not only words directly expressing public works but also relevant images and the names of political parties and politicians. Thus, perceptions of public works were widely analyzed. Their analysis revealed a change in people's impression of public works when the urgency and necessity of public construction projects were highlighted in the press. In Japan, the introduction of casinos is expected to have positive economic effects [7]. However, many people still have negative impressions of casinos. Gambling has been associated with suicide, unemployment, and productivity declines [8]. When making decisions, people choose the most satisfying option based on the available information [9].

In the context of casinos, decisions seem to be based on both positive and negative factors such as economic impact and social cost. Decision-making methods have also been explored. Of several factors affecting perceptions of casinos, Yan and Chee [10] clarified those that were significant using both the Analytic Hierarchy Process (AHP) and the conjoint analysis; the latter shows that practical decisions are made by combining several factors and is thus more useful than AHP, which simply compares factors. The cited authors compared the two methods, but not in the context of improving casino perceptions or marketing methods. Thus, although our methods are similar to theirs, the works differ in terms of perspective and significance. Conjoint analysis is not confined to casino perceptions; it is used to identify the most important factors involved in choosing a product when multiple possible factors may be in play [11, 12].

One study highlighted the influences of elements of the casino environment (building structure, degree of congestion, sound environment, etc.) on the mental state of people with problematic gambling habits using conjoint analysis [13]. Consumers are trying to select the more desirable option when choosing the most preferable product or service from among several [14, 15]. In such cases, consumers may use multiattribute decision making [16]. However, research that considers marketing elements among the constituent factors pertinent to casinos is limited, even considering research addressing environmental factors such as the structure of the facility [13]. Also, no research has addressed the question of how people understand words related to casinos. Therefore, we began by examining words related to casinos; we extracted such words using conjoint analysis.

Above, we focused principally on analytical methods. The reason why we used conjoint analysis will be given in the Methods section. Japanese casinos have been but poorly researched. Indeed, no keywords identifying casino preferences have been identified; our work is thus meaningful.

3. Experimental Details

3.1. Purpose. In this study, to determine the degree of desirability (hereinafter referred to as "preference") of various levels of casino, a conjoint analysis was conducted on the data obtained from a questionnaire survey. In this section, an outline of the survey and methods for deriving results will be described.

3.2. Method. Why did we use conjoint analysis to search for words reflecting casino preferences? The AHP mentioned above has much in common with conjoint analysis; both approaches preweight evaluation items (factors). However, we thought it inappropriate and unrealistic to weight combinations of individual evaluation items. When selecting products or services, customers make decisions after comprehensive evaluation of multiple factors at different levels. Additionally, the use of AHP was inappropriate because AHP preweights evaluation items; Japan does not yet have casinos and the casino-associated views of Japanese people remain completely unknown. Thus, it would be premature to use AHP in the present work. We concluded that conjoint analysis, which explicitly derives the influences of various factors, would be appropriate. We performed the following three steps:

(i) Step 1: We derived evaluation scores for all cards presented to subjects.

(ii) Step 2: We derived preferences for each level based on these scores.

(iii) Step 3: We assessed the influence of each factor based on the preferences for these levels.

In terms of Step 1, the evaluation scores of nine cards have already been derived [17]. The present research commences at Step 2. Clarification of preferences at all levels allows interactions among factors to be studied. Step 3 is currently underway.

3.3. Conjoint Analysis. During conjoint analysis, it is necessary to identify casino-specific factors and the levels thereof. First the setting in which factors are derived must be chosen. We extracted words pertaining to casinos and identified four factors based on a marketing mix framework (see Table 1).

In this paper, four factors were considered based on Kotler's marketing mix [18]. The Kotler marketing mix [18] is often used to identify factors affecting product purchase or use of a service. Takeuchi et al. [23] derived a marketing mix to improve service. Text mining was applied in this context.

In this paper, three levels, corresponding to choices within each of the mentioned factors in Table 1, were set. The three levels (casino-related words) for each factor are listed in Table 2. In terms of setting the levels, we referred to an overseas report [2] on casinos commissioned by the Cabinet Secretariat.

Figure 1 shows the four factors and twelve levels (three per factor) used to describe characteristics of casinos.

Nine conjoint cards were prepared by combining the twelve levels shown in Figure 1 based on the design of experiments (DoE) [24]. Each card shows a virtual casino created by combining four of the twelve levels. In this paper, the number of cards was decided with consideration given to minimizing the burden placed on participants answering the questionnaire. Therefore, the L9 orthogonal table was used, resulting in nine cards. Table 3 shows the nine conjoint cards.

Table 1: Extraction of words expressing factors based on the four Ps.

Factor	Extraction protocol
(Place): Location	"Place" of the casino means its location. Here, the relevant term is "Location."
(Price): Restrictions on betting	Words pertaining to money are extracted and regarded as factors corresponding to "Price." Casino money matters include (principally) entrance fees and bets. Entrance fees are controlled by Japanese governmental regulations; visitors cannot choose how much to pay. Also, the casino must charge only the legal fee. Therefore, we targeted words pertaining to betting. The word "Price" can be replaced by "The size of bets." It is necessary to keep in mind that the specific rewards vary by the games played and one's success rate. Thus, it is difficult to extract level-specific words. Many subjects will not know the situation that pertains if a specific level of betting is allowed. Subjects may choose their bets freely.
(Promotion): Atmosphere	This refers to promotion within the marketing mix. The purpose of promotion is to send a message to a target customer that increases the recognition of a product or service, improves its public image, and enhances sales [18]. It is indisputable that promotion directly influences consumption; moreover, promotion can also improve brand equity [19]. Promotion should not be viewed as a short-term approach to increasing sales. Indeed, promotional efforts should enhance brand equity in the long-term [20]. For casinos, brand equity involves awareness and a positive societal image. Given our focus on brand equity, we identified "atmosphere" as being important in terms of casino image.
(Product): Operating organization	As a casino is a single product, we extracted words reflecting reliability and quality. The characteristics of a product include safety, reliability, and suitability. When choosing electrical appliances and clothes, a consumer may focus on the country of origin and the manufacturer; these represent reliability [21]. The organization managing the casino is analogous to a manufacturer. The type of organization will influence whether a prospective customer would prefer to go to a casino or elsewhere.

Table 2: Extraction protocol for words expressing levels.

Factor	Level	Extraction of levels
Location	Odaiba Shinjuku Ginza	How will location affect how local residents value the casino? We extracted words at the level of "Local." All subjects commuted to either school or work; therefore, three distinct words with different characteristics served as the levels. All subjects were familiar with the features of each place. Odaiba, which is coastal, features an exhibition hall, a hotel, and amusement facilities. Many foreign tourists visit; land is available. Shinjuku is a business district in which the Tokyo Metropolitan Government Office and many other office buildings are located. This is one of the largest entertainment districts in Japan; many people come and go around the clock. There are many restaurants, bars, and entertainment facilities. Ginza has many old department stores and modern high-end stores. Although the luxury shopping street still has an old-fashioned atmosphere, many new commercial buildings have been constructed.
Restrictions on betting	Upper limit Without limit Lower limit	We nest set the general rules of Betting. We extracted words equivalent to "Restrictions" when betting money and also words equivalent to "Without limit," indicating that the amount of money that could be bet is not controlled. The word "lower limit" was extracted based on features of overseas casinos [2].
Atmosphere	Luxury Extraordinary Entertaining	For several reasons, we extracted words corresponding to levels of atmosphere, as follows. "Luxury" was extracted given that a casino requires a great deal of investment and commitment [22]. We extracted "extraordinary" because casinos are new ventures in Japan. We extracted "entertaining" because casino visitors can enjoy games and shows.
Operating organization	Domestic companies Overseas companies Local government	Domestic companies were extracted because they seek to enter the casino business, authorized by the IR bill. Overseas companies were extracted if they currently operate casinos overseas. Local government was extracted because it already manages public gambling in Japan.

Table 3: The nine conjoint cards.

Card No.	Location	Restrictions on betting	Atmosphere	Operating organization
1	Odaiba	Upper limit	Entertaining	Local government
2	Odaiba	Without limit	Luxury	Overseas companies
3	Odaiba	Lower limit	Extraordinary	Domestic companies
4	Shinjuku	Upper limit	Luxury	Domestic companies
5	Shinjuku	Without limit	Extraordinary	Local government
6	Shinjuku	Lower limit	Entertaining	Overseas companies
7	Ginza	Upper limit	Extraordinary	Overseas companies
8	Ginza	Without limit	Entertaining	Domestic companies
9	Ginza	Lower limit	Luxury	Local government

FACTORS	LEVELS
Location	Odaiba
	Shinjuku
	Ginza
Restrictions on Betting	Upper limit
	Without limit
	Lower limit
Atmosphere	Luxury
	Extraordinary
	Entertaining
Operating Organization	Domestic companies
	Overseas companies
	Local government

Figure 1: Factors and levels.

3.4. Survey. A web questionnaire featuring these conjoint cards was completed by 97 participants (71 males and 26 females). Twenty of the 97 reported that they had visited casinos in other countries. The study period ran for 34 days, from October 9 to November 12, 2017. A web questionnaire tool termed "Creative Survey" (https://creativesurvey.com/) was used. The target subjects were those who might visit casinos in future, particularly university undergraduate and graduate students. 20 of the 97 respondents reported experience with casinos in other countries. Thus, we focused on students, social workers, and students who were also working (all over 20 years of age).

The questionnaire survey was designed using a pairwise comparison method [25] that compared each of the nine cards shown in Table 2 with a set of two cards and then evaluated each card. The image shown in Figure 2 is a part of the screen shown as participants answered the questionnaire survey.

Subjects moved the cursor shown at the bottom of the screen to indicate which virtual casino described on the two cards (see Figure 2) was preferable. This procedure was

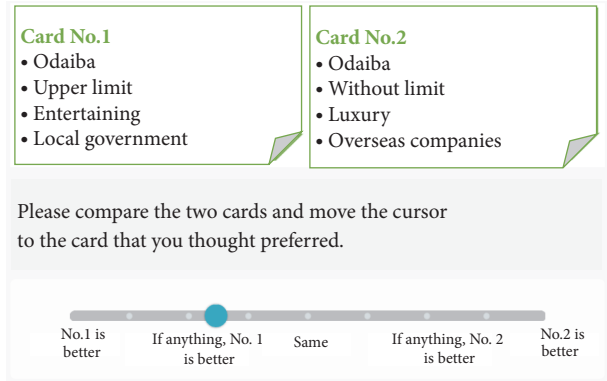


Figure 2: A sample questionnaire screen.

followed for all nine cards. Using the questionnaire survey, pairwise comparisons for each card were performed 36 times by each participant.

3.5. Determining Level Preferences for the Virtual Casino by Pairwise Comparison. The definitions of symbols and equations are as in Table 4.

The evaluation scores for all cards were calculated using

$$\sum_{N=1}^{97} E_N(\alpha) \tag{1}$$

Here, (1) is expressed as

$$E_N(\alpha) = \sum_{i=1}^{9}\sum_{j=1}^{9}\left(\left(0.5 - d_{ij}\right)u(C_i) + \left(d_{ij} - 0.5\right)v\left(C_j\right)\right) \tag{2}$$

The evaluation values of conjoint cards obtained from the 97 subjects were calculated using (1) and (2), where

$$u(C_i) = 1 \quad \text{if } 0.5 - d_{ij} > 0$$
$$\text{otherwise } 0$$
$$v\left(C_j\right) = 1 \quad \text{if } d_{ij} - 0.5 > 0 \tag{3}$$
$$\text{otherwise } 0$$

Table 4: The definitions of symbols.

Symbols	Definition
C_i, C_j	Card i · j(i<j, 1≤i≤9, 1≤j≤9)
C_{ij}	The combination of C_i and C_j
d_{ij}	The psychological value of C_{ij} $(0 \le d_{ij} \le 1)$
N	Number of general subjects
m	Number of factors (1≤m≤4)
n	Number of levels (1≤n≤12)
$\boldsymbol{\alpha}$	Score of C_i or C_j in C_{ij}
K	Number of pairwise comparisons (1≤k≤36)
X	Sample space
x_k	One sample space ($x_k \in C_{ij}$)
W	The combination of factor and level selected in X ($w_{mn} \in w(x_k)$)
Z_m	Preference within each level for the factor

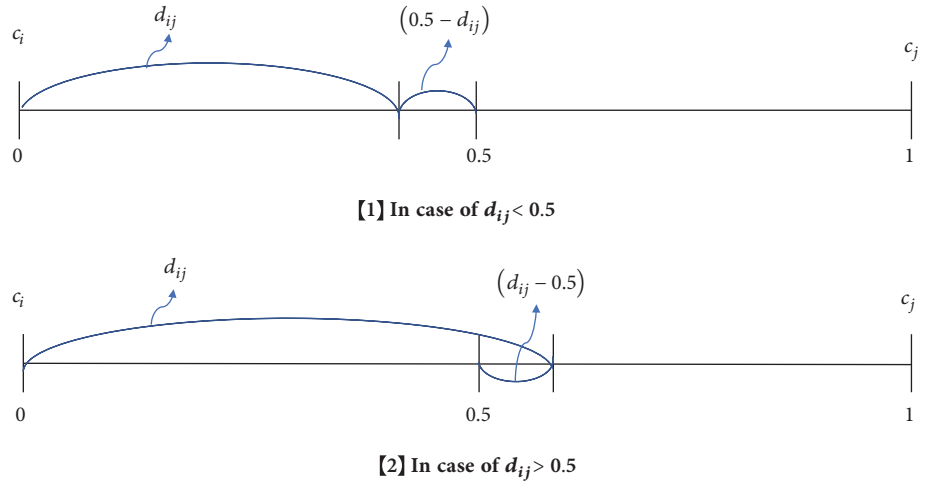


Figure 3: Process for evaluating each card by pairwise comparison.

$u(C_i)$ and $v(C_j)$ are classification functions that add scores to either C_i or C_j depending on the evaluation. The function can be described as follows (Figure 2 shows the questionnaire screen). In Figure 2, a subject moves the cursor from the center of the straight line ($d_{ij} = 0.5$) to the preferred side. For example, suppose that the subject considers Card C_j more favorable than Card C_i. In such a case, the cursor is moved from the center (same weights) to the right. Then, the d_{ij} value is the score of C_j. At this time C_i will not earn any points. Therefore, $d_{ij} = 0.5$ serves as the standard (Figure 3). We thus derived $E_N(\alpha)$.

If C_i and C_j fit into the sample space X by the *L9* orthogonal table based on the experimental design method, then one of them is $x_k \in C_{ij}$. Here, the combination of factors/levels that C_i and C_j can take is

$$\mathrm{W} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & w_{3n} \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{pmatrix} = w_{mn} \quad (0 \text{ or } 1), \tag{4}$$

and the preference for each level of the factor is shown below:

$$Z(x_k, w_{mn}) = \sum_{k=1}^{36} \sum_{m=1}^{4} \sum_{n=1}^{3} r w_{mn}(x_k) \tag{5}$$

In (5), "r" is a constant ($r > 0, r \in N$).

4. Results

Table 5 and Figure 4 show the results of conjoint analysis of the data obtained from the questionnaire. Ones' preference of all levels is depicted in both Table 5 and Figure 4 in a different way of expression. It is to be noted that the detailed values of ones' preference level calculated by (1), (2), and (5) are shown in Table 5, while the values of ones' preference level that are sorted in descending order is shown in Figure 4. It is supported to grasp at a glance on which level is the most popular among all the levels or which level is unpopular in Figure 4, which might be also referred to ones' preference order of the levels in the following chapter of discussion. Here, the preference for a given level of the virtual casino on each card was calculated in accordance with the

Table 5: Total level preferences for each factor.

Odaiba	Shinjuku	Ginza
338.4686	271.3594	252.3474
Upper limit	**Without limit**	**Lower limit**
330.7007	284.4942	241.9833
Luxury	**Extraordinary**	**Entertaining**
315.3692	296.2845	259.6396
Domestic companies	**Overseas companies**	**Local government**
311.3518	295.8827	264.0589

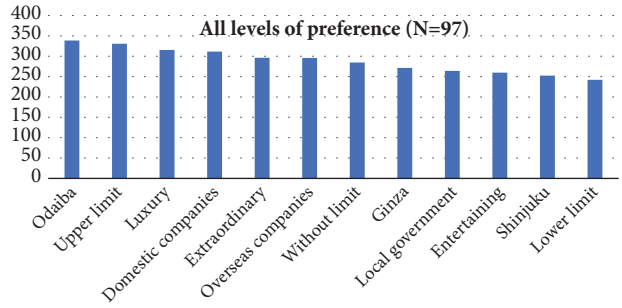


Figure 4: All levels of preference (N = 97).

procedure described in the fifth section of the third chapter (Experimental details; Determining level preferences for the virtual casino by pairwise comparison).

We now briefly summarize the results. Of the 12 levels, "Odaiba" was most favored. The, preferences for Shinjuku and Ginza were lower (Figure 4). "Location" significantly affected the level preferences as shown by the large variations in preferences among factors. Thus, location would greatly affect the decision to visit a casino. Also, levels allowing facile visualization of what is to be expected (such as "upper limit" and "luxury") were strongly preferred. We will discuss this in detail in the next chapter.

5. Discussion

This section addresses the question of how words expressing the level described on a highly rated card (i.e., the concept of a casino preferred by people) influenced casino preferences. Card X, which is thought to be the most preferred among the nine cards, was prepared by extracting the highest preference level for each factor one by one (see Figure 4). To assess the preference for all cards covering each of the three levels with respect to the four factors, 81 (= 34) cards must be compared one by one. However, in this paper, as cards were prepared based on DoE, the preference for all levels was assessed with only nine cards [24]. Therefore, Card X, which was not shown in the questionnaire, was prepared in this section. Figure 5 shows the procedure for preparing Card X, which incorporated the most preferred levels for each of the factors.

Furthermore, we extracted the lowest preference level of each factor to create Card Y. Figure 5 shows how we extracted highly preferred factors from each level. Using the same procedure, we extracted the least preferable level for each factor. Figure 6 shows card Y[1].

Based on cards X and Y described in Figures 5 and 6, we now focus on three points. First, we compare X, which features the highest standards for all factors, and card 2[1].

We next discuss card X per se, created on the basis of our results (see fourth chapter; the Result and Figure 5). We explain why four standards (Odaiba, upper limit, luxury, and domestic companies) were preferred; we refer to the nature of human decision-making described in the second section of this chapter.

Finally, to support the decision-making of subjects mentioned in the second section of this chapter, two new pairs of cards are compared with the aid of card Y. We then discuss words that should be employed in casino marketing strategies.

5.1. A Comparison between Two Cards with High Ratings. Figure 7 shows a comparison between Card X, configured with high preference levels, and Card No. 2, which had the highest evaluation rank in a previous study [17]. Calculation of the evaluation score for Card No. 2 was based on (1) and (2) and Figure 3 in the fifth section of third chapter (Experimental Details).

First, when comparing each level between the two cards shown in Figure 7, the levels of the location/atmosphere factors are common to both cards, i.e., Odaiba/luxury. Regarding the location, Odaiba is located on the waterfront alongside features of the coastal area, such as exhibition halls and large accommodation facilities. Therefore, Odaiba is preferred because it appears to be a suitable place for a casino. As a result, a positive impression, as expressed in language such as "gorgeous" and "similar to an overseas resort including a casino," is implied by the phrase "a casino in Odaiba" and it is therefore thought to be preferable.

Regarding the atmosphere, we considered that many participants judged that a feeling of luxury, related to the casino's roots [22], was the most preferable atmosphere. Even without knowing the historical background of a particular casino, the impression "casino = luxury" may have been created among participants. The word "luxury" appeared to imply a strong positive impression when compared to the other levels. For example, if the word "extraordinary" space gave the impression of a facility that small children would enjoy, such as the Tokyo Disney Resort, this may have been regarded as incompatible with a casino. The word "entertaining" was also seen as potentially carrying a wide

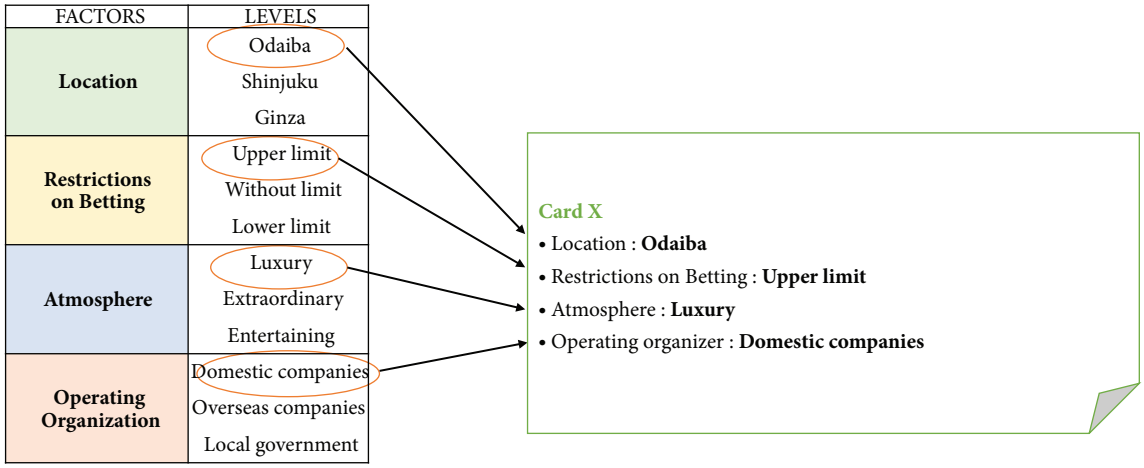


Figure 5: Preference for each factor based on the highest preference scores (Card X).

Figure 6: Preference for each factor based on the lowest preference scores (Card Y).

range of meanings. If the word "entertaining" were linked to the casino, the resulting impression may have been a negative one, associated with a place that is noisy or not elegant. Based on these results, it is speculated that it would be necessary to design a casino that offered a high-class atmosphere and high-quality hospitality if it were to become a preferred casino with a sense of luxury in Japan. In addition, it would be necessary to attract overseas luxury hotel chains to Japan and to share service know-how so as to provide high-quality customer service and a high level of service generally. This would make it possible to develop management strategies for the casino that would maintain a feeling of luxury. In any case, these considerations suggest that, by designing "a special place with a high-class feeling," such a casino would be preferred on the part of many Japanese people.

Second, these results indicated preferences for operating organizers who were well established as private enterprises, whereas casinos operated by local governments were not preferred. We considered that local governments were not favored as operating organizers because the pairing of the word "casino" with "local government" might invoke negative impressions of existing public gambling operations in Japan. In fact, some have pointed out that profits derived from public sporting events in Japan have peaked and are now in decline; these events no longer make major contributions to the exchequer [26]. Recently, many local governments have expressed concern about their financial situations; some are already bankrupt. In other words, Japanese public sports, originally viewed as social events enriching municipalities, are now driving municipalities into deficit. Thus, municipalities may not be trusted to run gambling facilities efficiently. In this context, casinos could be regarded negatively due to their association with images of existing public gambling features such as aging facilities, deficits, and gambling dependence. Thus, the results indicate that, to operate the casino most preferred by people in Japan, it will be necessary for companies that already operate overseas casinos to cooperate with companies aggressively entering this new market in Japan. Specifically, it will be necessary to formulate alliances between casinos operated in Japan and overseas casinos and to create a network of cooperation with casinos already in operation abroad. In this way, it would be possible to connect Japanese casinos to the peace of mind and pleasure of playing found in overseas casinos.

Finally, consideration was given to restrictions on betting, focusing on the presence or absence of an upper limit. Regarding restrictions on betting, preferences varied widely among participants. However, it can be inferred that language indicating that there is an upper limit may have evoked a sense of security. The following is an example of how a visitor might participate in gambling while maintaining a certain sense of security. When the casino has an upper limit, this limit indicates the amount of funds owned by the casino. In some cases, such a limit may increase anxiety about betting using IT technology, if, for example a comprehensible warning about excessive betting is issued. Moreover, although betting is not restricted, IT is now applied in casinos. For example, RFID tags have been inserted into genuine chips to distinguish them from fakes at the cashier's desk [27]. Casino visitors will appreciate such security, realizing that casinos deal with large amounts of money and must implement strict security checks. Even non-visitors may feel that IT usage renders the casino safer; perhaps they will then visit. Alternatively, it may be reasonable to consider a mechanism that restricts betting based on the participant's own financial margins and betting habits. For example, the upper limit might be divided into several stages, and the return can be considered based on an amount commensurate with

Figure 7: Comparison between two cards with high rating scores.

the upper limit. In any case, the word "upper limit" as an expression of restrictions on betting probably increases people's preference for a casino.

5.2. Consideration of the Four Most Favorable Levels. People make decisions within the framework of limited rationality [28]. People seek to be maximally rational, but their capacity to consider everything relevant may be compromised. Therefore, all decisions are partly irrational [29]. The lack of rationality is explained by the availability heuristic [30] and the representative heuristic [31]. The availability heuristic, in particular, features recall of familiar matters (because of physical closeness or via reports), impacting decision-making, as is evident in the preferences that we found. Card X reported the following data: location: Odaiba, restrictions on betting: upper limit, atmosphere: luxury, and operating organization: domestic companies. For each factor, it appeared that these four levels were selected because the availability heuristic was in play when subjects expressed their preferences.

Odaiba was the preferred location. In Japan, the casino bill was submitted to the Diet in 2013, and the government then implemented the "Improvement of the Base for International Tourism in a Near-Tokyo Seaside Sub-City (Odaiba Area)." In addition, some media reported that very large Japanese real estate enterprises and construction companies would participate in the project, creating the Odaiba Casino [32]. Although the bill introduced in 2013 was later abandoned, many news reports on the bill raised the level of public interest. Simply put, the idea that "Japanese mega-companies will build casinos in Odaiba" entered peoples' minds. Then, based on the availability heuristic, which commences with familiar objects, it became more likely that location: Odaiba and operating organization: domestic companies would be chosen.

The same pattern can be seen in terms of betting restrictions and "atmosphere". An upper limit on betting was favored. Many may be of the view that betting restrictions would prevent continuous betting. If gambling were in fact to be compulsorily halted, gamblers cannot continue. However, what if the maximum bet is so high that it is seldom attained? Perhaps some gamblers will continue because they simply have not reached the limit. Thus, a lower limit may be preferable; this would attract tourists who want to try some games just for fun. Therefore, the betting system will depend on what individuals actually want. Imposing a bet limit was favored partly because of frequent reports on casino rules overseas, such as limitations on entry and maximum withdrawals from teller machines [2]. Rather than winning percentages, or how much must be spent to improve the chances of winning, regulations [31] and standards described using familiar words [30] were the foci of attention of most subjects.

5.3. Supporting the Level Evaluation Process. In the first section of this chapter, we compared two cards featuring high preference levels. In addition, in the second section of this chapter, we discussed why four levels (Odaiba, Upper limit, Luxury, and Domestic companies) were highly preferred. We mentioned that a natural heuristic was in play. Here, we compare two new pairs of cards to amplify the discussion of the previous chapter. We first compare cards No. 6^2 and card Y. Figure 6 refers to card Y.

We next compare cards X and Y, which were created by reference to the level selection order. When we compare these two pairs, we discuss the optimal words of casino marketing proposals.

Figure 8 compares card No. 6, which was least evaluate and card Y (featuring low preference levels).

The cards shown in Figure 8 are identical except for the operating organization. This shows that casino evaluations were made at three levels, Shinjuku, entertainment, and the lower limit of betting. As Table 2 indicated earlier, Shinjuku has two aspects; safety is of concern because Shinjuku has a large red-light district. Also, an illegal casino discovered in Shinjuku was reported in the media; thus, the phrase "Casino located in Shinjuku" might have negative connotations. Also, the combination "Shinjuku" and "entertaining" may have created negative images. Of course, Shinjuku is not the only place where public safety is of concern. However, Shinjuku is a famous red-light district, and illegal casinos and restaurants are sometimes reported in the media. Thus, the combinations of words on cards Y and No. 6 created a poor impression. We expect that subjects associated illegal casinos with these cards and that availability heuristics [30] affected the preferences.

Cards No. 2 and X were in terms of desirable (Figure 7) and unwanted casinos. Cards No. 6 and X have also been compared (Figure 8). Here we compare cards X and Y, given our outcomes. By comparing conflicting cards, we can discover how society views casinos. Figure 9 shows cards X and Y.

As shown in Figure 9, cards X and Y differ markedly in their combinations of levels.

Card X features the most favorable levels and Card Y the least favorable levels. Heuristics markedly influence

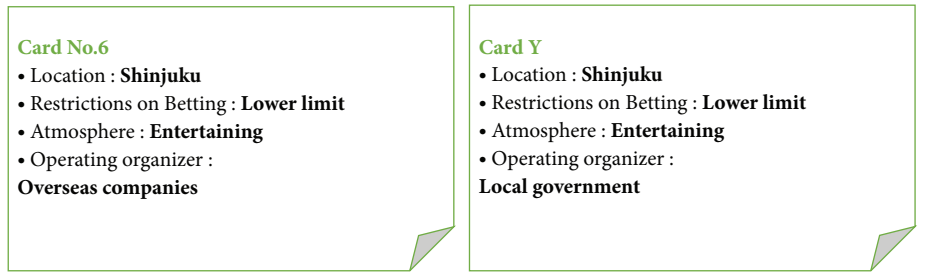


Figure 8: A comparison between two cards with low rating scores.

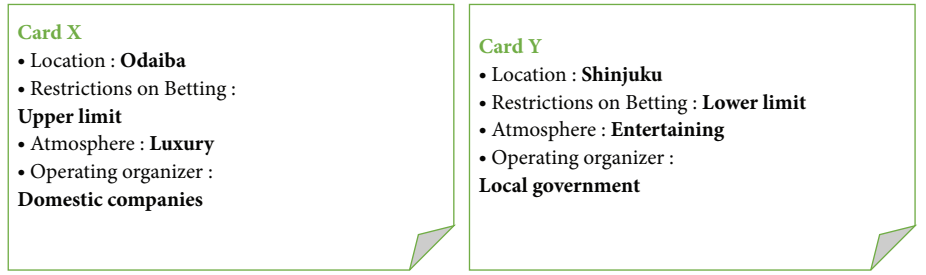


Figure 9: A comparison of two conflicting cards.

evaluation, regardless of the cards per se. Decision making may thus be based on heuristics. As shown in Figure 9, the words on Card X were familiar casino-related words. Especially, "Luxury," which described the atmosphere, imparted a positive impression. In contrast, we assume that the words on Card Y were less acceptable, given earlier widely reported negative incidents. On Card Y, furthermore, the combination of "entertaining" and "Shinjuku" created an image of an unsafe red-light district. Thus, even words that do not express emotions (such as place words and words imparting restrictions) in fact create emotions, depending on their combinations. Therefore, we conclude that outcomes will differ even among evaluations based on the same heuristics. If so, it is necessary, initially, to increase casino awareness in Japan by using familiar casino-related words that subjects associate with positive feelings.

As shown in Figure 9, "Atmosphere" was readily reflected by "Luxury", given the history of casinos [22]. Thus, it will be possible to improve casino recognition in Japan by optimizing the human decision-making process, by familiarizing people with casino-related words. Casinos will gradually become more accepted, as will their business enterprises (well-known casino organizers); media reports will increasingly use words emphasizing the high standards to which casinos are held. Matsui [33] studied the interactions between words and marketing. Marketing generates new words; the spread of trending keywords enhances marketing. More particularly, "keyword of generation" created by a marketer can logically fill the needs of consumers, eventually strengthening the new word-based trend. Japan currently lacks casinos, and few people are familiar with overseas casinos. It is important to improve casino recognition; marketers must create the trend by using terms associated with casinos in an environment where people will be likely to hear those terms.

6. Conclusion

In this paper, 97 participants evaluated virtual casinos using conjoint cards, providing information on their preferences in relation to various characteristics of casinos. In addition, we considered qualities associated with high levels of preference and the manner in which words depicting those qualities may have influenced participants' impressions of the casino.

This study has several limitations. First, only 97 participants were included, making it a small-scale survey. The participant characteristics were also limited, and it is understood that the results obtained from this questionnaire survey are only applicable to a very limited situation. The numbers of factors and their levels were also limited to reduce the burden on the participants, so not all casino and IR facilities were covered. In the future, it will be necessary to consider ways to evaluate further people's impressions of casinos by setting more factors and levels and, thereby, evaluating more words related to casinos.

Data Availability

The .xlsx data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to express their sincere gratitude to all 97 individuals who participated in the survey.

Endnotes

1. Card No. 2 of the nine cards received the most evaluations [17].
2. Card No. 6 of the nine cards received the least evaluations [17].

References

[1] K. Kitagawa, T. Naito, and K. Terada, "POlysemy of light in text description of buildings," *Journal of Architecture and Planning (Transactions of AIJ)*, vol. 77, no. 680, pp. 2345–2353, 2012.

[2] KPMG AZSA LLC, Overseas case study report on Specified Integrated Resort Facilities, 2015.

[3] Y. Masuko, "Structural change in the Macau casino industry -In reference of open-door policy to unvested capital from abroad," *Journal of Japan Society for Global Social and Cultural studies*, vol. 13, no. 1, pp. 26–36, 2016.

[4] I. Tanioka, *Attitude toward the legalization of casino gaming in Japan: The difference in regions, urbanization levels, and by socio-economic background*, vol. 3, Institute of Social Science University of Tokyo, 2014.

[5] Japan Productivity Center, *Leisure Hakusho 2017*, vol. 2, 2017.

[6] K. Tanaka and Y. Kanda, "Reserch of people's image changing towards public works," *Journal of Japan Society of Civil Engineers, Ser. F4 (Construction and Management)*, vol. 69, no. 4, pp. 1–7, 2013.

[7] A. Daswani and G. Choi, "Global gaming rising sun to outshine Vegas: Japan set to launch casinos," *Citi Research Global Gaming*, pp. 3–35, 2013.

[8] E. L. Grinols, *Gambling in America: Costs and Benefits*, CBG: Cambridge University Press, 2016.

[9] H. A. Simon, *Administrative behavior: A study of the decision-making process in administrative organization*, The Free Press, New York, NY, USA, 1997.

[10] K. H. Yen and K. L. Che, "Empirical examination of AHP and conjoint analysis on casino attributes in Macau," in *Proceedings of an International Conference on Public Welfare and Gaming Industry, Beijing*, pp. 327–350, 2010.

[11] K. Fujiwara, A. Yagahara, G. Inoue, T. Kitagawa, and K. Ogasawara, "Investigation of Preferences in Working Environment of Radiological Technologists Using Conjoint Analysis," *Japanese Journal of Radiological Technology*, vol. 73, no. 8, pp. 626–635, 2017.

[12] M. Matsushita, K. Harada, and T. Arao, "Incentive program to strengthen motivation for increasing physical activity via conjoint analysis," *Japanese Journal of Public Health*, vol. 64, no. 4, pp. 197–206, 2017.

[13] K. F. Gough, "The influence of casino architecture and structure on problem gambling behavior," in *Proceedings of the ECRM2015-Proceedings of the 14th European Conference on Research Methods for Business and Management Studies*, pp. 66–73, 2015.

[14] S. Iyengar, *The art of choosing*, Grand Central Publishing, New York, NY, USA, 2011.

[15] T. Kan, *Excel ensyu de manabu tahennryou kaiseki - kaiki bunnsekihannbetsu bunnsekiconjoint bunnseki henn-[Multivariate analysis learned by Excel exercises-Regression analysis-discriminant analysis-conjoint analysis] , Tokyo: Ohmsha*, Ohmsha, Tokyo, 2016.

[16] K. Takemura, R. Higurashi, and Y. Tamari, "Cognitive effort accuracy of decision strategies in multi-attribute decision-making process: A behavioral decision theoretic approach using computer simulation technique," *Cognitive Studies*, vol. 22, no. 3, pp. 368–387, 2015.

[17] N. Komiya and J. Nakamura, "A study on cognitive improvement and evaluation of casinos in Japan using conjoint cards," *Technical Report of IEICE*, vol. 117, no. 440, pp. 21–26, 2018.

[18] P. Kotler, *Kotler on Marketing: How to create, win and dominate markets*, The Free Press, New York, NY, USA, 1999.

[19] D. A. Arker, *Managing Brand Equity*, The Free Press, New York, NY, USA, 1991.

[20] T. Kobayashi, "Brand-based marketing: the effect of a hidden marketing system," *The Business Review*, vol. 49, no. 4, pp. 113–133, 1999.

[21] K. I. Al-Sulaiti and M. J. Baker, "Country of origin effects: A literature review," *Marketing Intelligence & Planning*, vol. 16, no. 3, pp. 150–199, 1998.

[22] J. R. Strauss, *Challenging Corporate Social Responsibility*, Routledge, 2015.

[23] H. Takeuchi, Y. Sugiyama, C. Ota, and T. Yamaguchi, "Marketing analysis using marketing mixture and text mining," in *Proceedings of the The 24th Annual Conference of the Japanese Society for Artificial Intelligence*, pp. 1–4, 2010.

[24] M. Sueyoshi, *Excel business toukei bunnseki[Excel business statistical analysis]*, Shyoeisya, Tokyo, 2017.

[25] H. Takagi, "Practical statistical tests: machine learning (3) Significance tests for human subjective tests Systems," *Control and Information*, vol. 58, no. 12, pp. 514–520, 2014.

[26] H. Fukui, "Formation and prospects of public competition: Focusing on motorboat races," *Hosei University Repository Public Policy and Social Governance*, vol. 5, pp. 149–163, 2017.

[27] M. Takahashi, "RFID Tag Antennas," *IEICE Communications Society Magazine*, vol. 2008, no. 7, pp. 7_51–7_58, 2008.

[28] H. A. Simon, *A Behavioral Model of Rational Choice*, Wiley, New York, NY, USA, 1957.

[29] M. H. Bazerman and D. A. Moore, *Judgment in Management Decision Making*, Wiley, New York, NY, USA, 2008.

[30] A. Tversky and D. Kahneman, "Availability: a heuristic for judging frequency and probability," *Cognitive Psychology*, vol. 5, no. 2, pp. 207–232, 1973.

[31] A. Tversky and D. Kahneman, "Judgment under uncertainty: heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[32] J. Fujita and N. Lane, "Mitsui-fudosan Fuji-television Kajima, Daiba de casino kensetsua teian Mitsui fudosan Fuji-television Kajima, casino construction proposal in Daiba," *Reuters*, 2013.

[33] T. Matsui, "Language and marketing: A text analysis of the dynamic change of the shared cognition on "healing" in Japan between 1998–2007," *Organizational Science*, vol. 46, no. 3, Article ID 19982007, pp. 87–99, 2013.

Hindawi
Advances in Human-Computer Interaction
Volume 2019, Article ID 4951387, 11 pages
https://doi.org/10.1155/2019/4951387

Research Article

Effect of Employees' Values on Employee Satisfaction in Japanese Retail and Service Industries

Tomonori Matsuki[1,2] **and Jun Nakamura**[1]

[1] *Shibaura Institute of Technology, Japan*
[2] *Recruit Management Solutions Co., Ltd., Japan*

Correspondence should be addressed to Tomonori Matsuki; na17105@shibaura-it.ac.jp

Received 29 June 2018; Accepted 17 October 2018; Published 1 January 2019

Guest Editor: Akinori Abe

The Japanese workforce has decreased rapidly over the past few decades, and this is expected to continue. Retail and service industries are already experiencing human-resource shortages. In these industries, nonregular employees feature prominently. For most companies, recruitment is difficult, and employees change jobs often, making securing staff an important business issue. Nonregular and regular employees are treated differently; the problem is thus partly social in nature. However, some nonregular employees are content, although their work conditions are not good. Here, text mining was used to explore differences between the values of regular and nonregular employees in the retail and service industries.

1. Introduction

The retail and service industry, which is labor intensive, is facing a turning point in human-resource management due to a decreasing labor force population and diminishing value for employees' work (the term "values" is defined as a way of thinking about work and workplace in this paper). In this industry, companies have used large numbers of nonregular employees for many years to accommodate fluctuations in supply and demand. According to MHLW [1], about 70% of nonregular employees in the accommodation and food service industries are nonregular employees. Urgent measures are required to identify how to grow the abilities and motivation of nonregular employees. Additionally, the high turnover rate of employees is a significant issue. Maintaining and training human resources have become increasingly demanding management tasks.

Many companies have legacy human-resource systems and employ students and housewives as part-time staff. These companies are not aligned with labor market needs.

Furthermore, the response to nonregular employees is often decided by the planning department at headquarters and does not take into account the values of nonregular employees recently engaged at stores.

Therefore, the author conducted a survey at seven major chains in the service industry to investigate intrinsic employee satisfaction (ES) factors and extrinsic ES factors such as the values of employees' work. The findings of this paper will provide guidance to store management in the service industry.

This paper is divided into four subsections. First, the current situation in the Japanese labor market is explained. Second, the effects of service industry characteristics on employees' values most affected by the labor market are described. Third, features of the retail and service industry are detailed. Finally, working conditions in the retail and service industry are described.

1.1. The Current Situation in the Japanese Labor Market. In recent years, the labor force in Japan has decreased sharply and this is expected to continue. The decrease is remarkably large compared to other developed countries (Figure 1). Thus, companies often seek to employ students, housewives, and the elderly, who comprise "nonregular" employees. The Ministry of Health, Labor and Welfare (MHLW) [1] published the "Vision for Working Habits", stating "Eschewing the bipolarized notion of regular and nonregular employment,

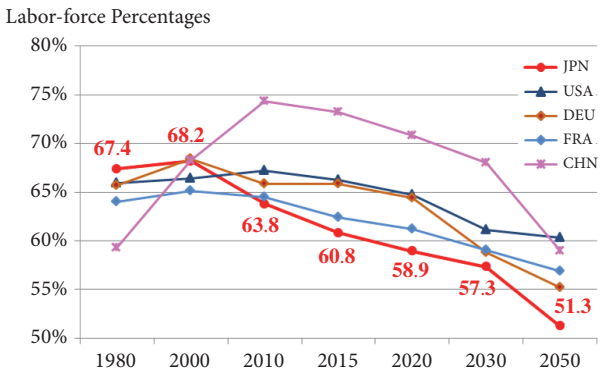


Figure 1: **Trends in labor force percentages by country**. The labor force percentage is the percentage of the total population aged 15–64. Source: Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, *World Population Prospects*, 2015 [2].

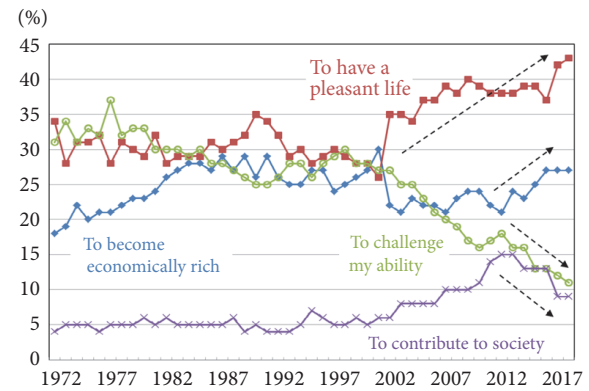


Figure 2: **Changing work attitudes of new graduates**. The vertical axis is the percentage of responses (n = 1882). Source: Japan Productivity Center, 2012.

the focus should be on stable employment, and improvements in worker morale and capability, thus improving and developing the Japanese economy and society."

A "nonregular employee" is difficult to define, as people select various modes of working. The MHLW defined "nonregular employees" as employees who work for a flexible period and are either not in full-time work or not in direct employment. The MHLW identified the problems of nonregular employment as instability, difficulty in achieving economic independence, inadequate career progression, and lack of a safety net.

1.2. Features of the Retail and Service Industries. Employees in the retail and service industries were investigated in terms of attitudes and satisfaction. These industries are very labor intensive, so a lack of workers greatly affects their business growth. In these industries, customer demands vary greatly by time of day and day of the week, and most companies have traditionally hired nonregular employees and most did not require professional qualifications. However, nonregular employees who are paid and treated poorly are often better performers than regular employees. For example, in a clothes shop, nonregular employees sometimes generated higher sales than regular employees. Also, in the food service chains, some nonregular employees have worked longer than regular employees and, consequently, have better operational skills.

1.3. Factors Affecting Employees' Values. This study analyzed the attitudes of nonregular and regular employees toward work. According to data in a report from the Japan Productivity Center [3], attitudes to work have changed recently. The report showed that the percentages of young people (under 25 years old) who "want to have a pleasant life" and "want to contribute to society" have risen since 2000, but the percentages who "want to challenge my ability" and "want to become economically rich" have declined (Figure 2). Thus, research into nonregular employees' attitudes may reveal some novel motivations.

1.4. Working Conditions of the Retail and Service Industries. Retail service is emotional labor [4], as employees must constantly control their emotions to accommodate demanding customers. Nonregular employees in accommodation and food service industries are paid at only 32.8% the rate of regular employees, which is the lowest of all industries (average 49.4%) [5, 6]. In addition, the percentage of full-time employees is low (30.8%; [1]), and employees frequently leave, rendering training and evaluation difficult. The turnover rate within 3 years is 50.2% for university and 64.4% for high-school graduates [5, 6], significantly higher than the average (32.2% for all industries [university graduates]; 40.8% [high-school graduates]). Traditional attitudes toward nonregular employees persist and management often has low expectations: "We do not allow these employees for much responsibility", "We do not expect high quality" and "Job turnover is high, but this cannot be helped" [7].

Quantitative correlations analysis and qualitative text mining [8] were used to define what satisfies nonregular employees. In this paper, the retail and service industries were targeted because they often do not require professional qualifications or business proficiency. Thus, management will understand the findings; text mining will not unearth technical terms. Also, in sales departments, regular and nonregular employees play similar roles, sharing many tasks. Therefore, it is not difficult to compare their respective views and values based on employment patterns.

1.5. Research Questions. Many studies have investigated the influence of extrinsic factors on ES, but few have focused on intrinsic factors. This is because intrinsic factors are difficult to measure quantitatively. However, elucidating the intrinsic factors is important to understanding employee motivation and to increasing the effectiveness of various measures to enhance performance. Therefore, this study analyzed intrinsic factors based on qualitative data and investigated whether intrinsic factors affected ES. In addition, the analysis was based on conditions specific to the service industry, which include differences in employment patterns. Therefore, the research questions in this paper were as follows:

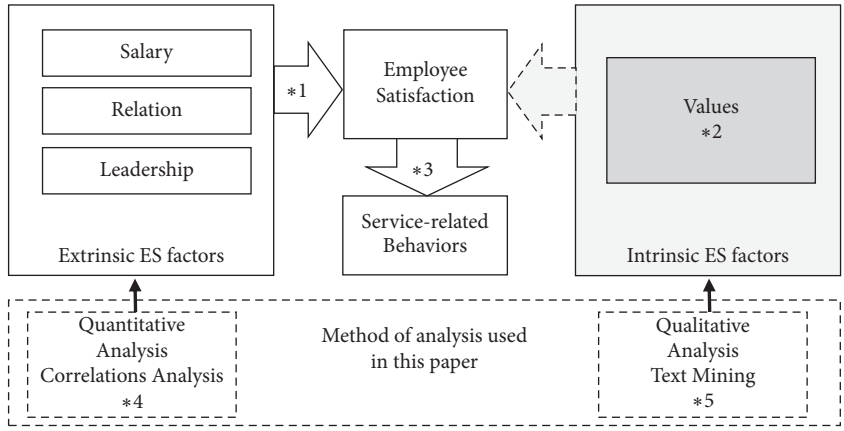


Figure 3: **Overview of the research**. Note: ( ) are the section numbers relating to previous research and method in this paper. ∗1 concerns the effects of extrinsic factors on ES (Section 2.1). ∗2 concerns the intrinsic ES factors (Section 2.2). ∗3 concerns the effects of ES on service-related behaviors (Section 2.3). ∗4 concerns the analytical method on extrinsic ES factors (Section 3.3). ∗5 concerns the analytical method on intrinsic ES factors (Section 3.4).

Research question 1: Do the attitudes of employees in the retail and service industries affect their satisfaction?

Research question 2: Do attitudes differ by work, workplace, and/or employment pattern?

It is thought that answering these two research questions will be helpful to both the service industry and companies that hire nonregular employees.

1.6. Hypothesis. In the retail and service industries, even under the same conditions, employee satisfaction, behaviors, and performance differ by individual. This implies that in considering the factors that impact employee satisfaction there are some that cannot be fully explained only by extrinsic factors. Therefore, two hypotheses are considered:

Hypothesis 1: Values influence employee satisfaction. Changing attitudes to, and values associated with, work and the workplace influence employee satisfaction.

Hypothesis 2: Values differ by employment pattern.

2. Previous Research

Figure 3 provides an overview of previous research. This study focused on intrinsic and extrinsic drivers of employee satisfaction (ES). Such drivers and their effects on store performance were examined (Figure 3).

There is no fixed definition of ES in the literature. For example, some studies have focused on personal aspects such as treatment and wages, while others deal with the quality of human relationships in the workplace.

In this paper, ES is defined as the overall satisfaction level of working in a shop. There are many nonregular employees working in shops in the service industry, and they typically work in limited spaces with few coworkers. This paper analyzes how employee values influence employee satisfaction in such an environment. Our definition of ES does not consider the degree of satisfaction related to position, honor, and future return. Herein, ES is only based on daily work. Employees cannot control external factors, although such factors influence ES.

2.1. Effects of Extrinsic Factors on Employee Satisfaction. This subsection outlines previous studies related to the influence of extrinsic ES factors on employee satisfaction (asterisk 1 in Figure 3). Alderfer [9] used a Likert questionnaire to investigate the motivations of 300 factory workers based on salary, benefits, supervisors, colleagues, and growth. In this paper, authors use the same method of questionnaire. Herzberg [10] evaluated 1,683 engineers, scientists, military personnel, and nursing staff in terms of their "satisfied" or "dissatisfied" status (positive motivations and negative hygiene factors, respectively). Harter et al. [11] meta-analyzed studies on 36 companies and 7,939 organizations (198,000 people in all) in terms of employee satisfaction (12 items). Authors referred these theories for structures of question items in ES factors. Carter and Bughurst [12] conducted focus group interviews on the importance of leadership to restaurant employees; the data have been used in the food service industry. Authors referred situation of service industries of the other country. Eggers and Kaul [13] analyzed the patenting patterns of various companies, exploring whether the inventions reflected high- or low-level motivation. It is interesting to show the relationship between ES and innovation. In this paper, By referring to research of other industry, authors can find characteristics of the retail and service industry. Pugliesi [14] noted that the effects of emotional labor reflect other work-related conditions. Indeed, emotionally laden work can sometimes have negative effects, including job stress and poor satisfaction.

It is important to consider both ratio centric and emotional work when analyzing the ES of retail and service industry employees. The relationships between employee motivation and behavior were reviewed above. However, few studies have focused based on employment. Given the

uniqueness of nonregular Japanese employees, it is important to analyze differences between regular and nonregular employees in terms of motivation, not simply money and treatment. Thus, prior studies established the model used in this paper. This study hypothesized that leadership, salary, and the ingenuity of management would influence employee behavior.

2.2. Intrinsic Factors on Employee Satisfaction. This subsection describes previous studies of values related to intrinsic ES factors (asterisk 2 in Figure 3). Schwarts [15] performed questionnaire surveys in 44 countries and classified universal human values into 10 types by motivational purpose. Due to factors base on personal types, the point that affects motivation is a reference to this paper. Ros [16] grouped the values as "intrinsic", "extrinsic", "social", or "prestige-related" and discussed their interrelationships. In this paper extrinsic factors and intrinsic factors are separately positioned as factors affecting ES. Twenge [17] explored differences in the morality and values of baby boomers and those born after 1982. To and Tam [18] explored the values attached to the work of migrant women in China, their employment compensation, their satisfaction, and differences in values. The difference in values by these generations is often a theme. How to analyze the difference by these attributes is helpful.

Such studies structured and classified working values, but it remains unclear whether values intrinsically influence employee satisfaction. Values differ among individuals and are difficult to quantify. Thus, it has been impossible to conduct a study showing the relationship with employee satisfaction. However, with the development of an effective tool to quantify qualitative information, it has become possible to analyze such values.

2.3. Effects of Employee Satisfaction on Service-Related Behaviors. This subsection describes previous studies related to the influence of employee satisfaction on employees (asterisk 3 in Figure 3). Sun et al. [19] found that service-related actions affected employee productivity, turnover, and ultimately corporate performance. Heskett et al. [20] emphasized that profitability, customer loyalty, and employee satisfaction constituted a "service profit chain". The theme in this paper is a factor influencing ES, and the behavior change is out of the scope. However, it will increase the significance of this research that ES has a big influence to achieve high performance. Amabile and Kramer [21] analyzed individual diaries, and observed events in the workplace and the inner workings of employees. Performance was influenced by recognition, emotion, and motivation associated with the workplace. Morrison [22] stated that, to avoid organizational problems, employees must receive information from those in higher positions; otherwise employees will be silent and fail to deliver important information to superiors. Matsuki and Nakamura [23] developed an ES–customer service (CS) model for the service industry using the service profit chain concept of Heskett [20] and the two-factor theory of Herzberg [10]. Their model identifies factors affecting employee behavior in terms of ES per se, CS, and store performance. Quantitative analysis of relationships between service behavior and performance is commonplace in both academia and the real world. However, it is often difficult to identify actions affecting CS or employee performance.

Thus, this study evaluated employees from several companies. The studies cited above showed that employment conditions and the workplace influence employee awareness and behavior, but intrinsic employee values may also affect ES. This study focused on intrinsic employee values.

There is no fixed definition of ES in the literature. ES has been investigated in the context of discrete personal aspects such as treatment or wages, and in other cases in relation to the quality of human relationships in the workplace. Herein, ES refers to the overall satisfaction level of working in a shop. There are many nonregular workers in shops in the retail and service industry, and they typically work in limited spaces with few coworkers. This paper analyzes how employee values influence employee satisfaction in this environment. Therefore, our ES is not the degree of satisfaction related to position, honor and future return but, rather, to the current situation in the shop. Additionally, in this paper, extrinsic factors are defined as not controllable by companies while intrinsic factors can be controlled. Thus, companies should recognize the values of employees more and provide them with better support.

3. Methods

A survey of employees in the retail and service industries was conducted to investigate effects on customer satisfaction, both quantitatively and qualitatively. The overall survey is shown in Figure 4.

3.1. Survey Planning. A survey of employees at seven major companies in the retail and service industries was conducted. The questionnaire is shown in Table 1. The questions consisted of 11 single-answer items for ES and its factors and a free-answer item proposed by the employees. Table 1 shows the data obtained.

3.2. Survey Implications. The survey was conducted from June 10 to October 25, 2016 using an anonymous, Web-based response system. Multiple responses from the same source (e.g., smartphone, tablet, or PC) were not permitted. Each employee entered a QR code. One manager of a food service company commented that many regular and nonregular employees completed the questionnaire during lunch breaks or on their way home.

In total, there were 2,513 ES responses and 653 free-answer proposals. The responses to the survey by employment pattern are shown in Table 2.

3.3. Quantitative Correlation Analysis. The authors investigated the relationship between extrinsic factors and employee satisfaction through correlation analysis as a preliminary research. Correlation coefficients between ES and ES factors were calculated for regular employees and nonregular employees, respectively, and significant differences determined. If there is no statistically significant difference between the two, then the effect of extrinsic factors is the

Table 1: Whole question items in the research.

Variables/Category	Question Items (English)	Question Method
Salary	I think this store salary and working conditions is good.	Single Answer 5 Level
Equipment Environment	I think facilities and environment of this store is good.	
Leadership	I think this store takes good leadership supervisor.	
Personal Growth of Employees	I think that work in this store leads you to grow.	
Creative task	I think we work many creative tasks as well as routine tasks.	
Customers' reactions	I think this store service are happy customers.	
Financial performance	I think this store's performance will be better again.	
Promptness of service	I think the employee's "rapid response" in this store.	
Improvements in service	I think employees corresponds with a smile in this store.	
Service w/smile	I think not manual, think for yourself, acting as employees of this store.	
Employee Satisfaction	I think I am satisfied with working at this store.	
Proposals for improving stores	Please feel free to have opinions and ideas for making your store better with innovations and improvements.	Free Answer open-ended question

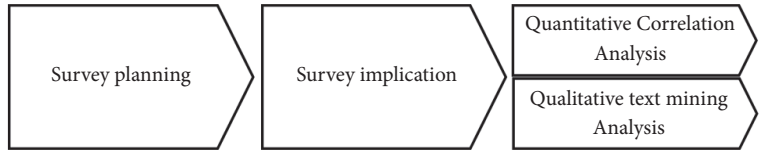


Figure 4: Overview of the research.

Table 2: Questions and numbers of respondents.

Regular employees	Non-regular employees	Total
773 (270)	1,740 (383)	2,513 (653)

Note: ( ) indicates respondents that answered the item "improvement proposals" using free text.

same for regular employees and nonregular employees. The difference between intrinsic factors then becomes clearer.

3.4. Qualitative Text Mining. The free-text comments were mined and analyzed using text analytics to quantify and visualize the comments.

3.4.1. Text Mining Software. Text mining was performed using the tool "Mieru-ka Engine" by Plus-Alpha Consulting. Text mining divides comments into words or phrases, and analyzes their frequencies and correlations. For example, Ford et al. [24] used text mining to extract items affecting employee engagement. Here, the employees' responses were analyzed both quantitatively and qualitatively in terms of ES and values, and whether differences were related to employment pattern was explored.

3.4.2. Statistical Dependency Analysis of Employee Satisfaction and Views. Statistical dependency analysis "visualizes the qualitative context inherent in text data". Such text mining analyzes qualitative language data, aggregates the associations and dependencies of words and phrases, and expresses relationships between words in diagrams. The comments of the employees were divided into those by employees who were (≥3 points) and were not (≤2 points) satisfied, and further divided into those made by regular and nonregular employees. This formed groups of regular employees who were or were not satisfied (Groups A and B, respectively), and similar groups of nonregular workers (Groups C and D, respectively). Statistical dependency diagrams were then created. It was assumed that differences in the keywords in the free comments used by employees with high ES levels and those with low-ES levels reflected differences in employee satisfaction. Differences in keyword use based on employment pattern were also analyzed. Sentences are separated into each word and count the number of the specified word and its connected word. When qualifying a specified word, draw an arrow toward the specified word and draw backward arrows if modification is made from the specified word.

3.4.3. Keyword Ranking Analysis. Keyword ranking was used to extract keywords representing values. At the end of the survey, questions regarding proposals for store improvement were provided in free-comment form. Using keyword rank analysis, the keyword appearance rates in the respondents' proposals were extracted.

In addition, the top 20 keywords used by regular and nonregular employees were displayed in descending order and the values of d and s were analyzed by employment pattern.

$$\text{Keyword appearance rate} = \frac{\text{Number of keyword users}}{\text{Free comment number of respondents}} \quad (1)$$

Table 3: Relationships between employee satisfaction and various subfactors.

ES factor (subfactors)	Regular employee	Non-regular employee	z values
(1) Wages and treatment	0.38	0.36	0.46
(2) Workload	0.31	0.31	1.00
(3) The environment	0.39	0.41	−0.69
(4) Relationships in the workplace	0.58	0.54	1.39
(5) Leadership of supervisor	0.46	0.48	−0.46
(6) Company policy and administration	0.53	0.53	1.00
(7) Responsibility	0.50	0.49	0.23
(8) Evaluation	0.51	0.54	−0.92
(9) Customer satisfaction	0.47	0.46	0.23
(10) Growth	0.67	0.67	1.00

Note: the values are correlation coefficients between 10 possible ES factors and CS. The z values are those revealing a significant difference between the two correlation coefficients. When $z > |1.96|$, significance was assumed at a 5% rejection probability.

Table 4: Comparison of four groups in keyword map.

Group	ES Score	Employment pattern	n	keywords
A	≧3	Regular employees	211	human resource, staff, shop, part-time, supervisor, employee, motivation, opinion, whole work, service, pit, time
B	≦2	Regular employees	562	human resource, staff, shop, part-time supervisor, necessity, opportunity, company, everyone, salary, status, conscienceless, communication, meeting
C	≧3	Non-regular employees	344	Supervisor, employee, staff, opinion shift, air conditioner, control, sales floor, direction, teaching
D	≦2	Non-regular employees	1,396	Brand, kitchen, day, hourly wage
Total			2,513	

4. Results

Regular and nonregular employees coexist under different conditions, but seem to share values relevant to work and the workplace. Regular and nonregular employees were compared in terms of the effects of 10 subfactors on satisfaction. However, no significant differences were found by employment pattern (Table 3).

4.1. Quantitative Correlation Analysis. Correlation coefficients between ES and 10 Extrinsic ES factors were calculated for regular employees and nonregular employees, respectively. Although each extrinsic ES factor has an effect on employee satisfaction, no statistical significance was found for the difference in employment pattern.

4.2. Comparison Keyword Map Analysis by ES and Employment Pattern. Keyword analysis was performed to extract keywords that are characteristic to each group. Differences between such keywords were interpreted. Unlike the problem of the selection formula, what is routinely conscious is expressed directly without being induced by some answers. For this reason, text mining was selected as the method to analyze intrinsic factors.

The keyword map is useful to visualize the entire picture of distinctive differences. By using attribute data, keywords characteristic to each group were compared in a map diagram. The blue circle indicates attribute data. The yellow circle linked to the blue circle shows keywords that are characteristically appearing in that attribute. The green circle encompasses common keywords. A yellow line connects a group to common keywords and the number of occurrences is shown in parentheses.

Satisfied employees (ES ≧3) used nouns relating to people such as "staff", "employee", or "superiors" more often than dissatisfied employees and the nouns reflecting improvements were more concrete. Regular employees commonly mentioned "time", "service", and "motivation"; while nonregular employees mentioned "shift" and "environment" (Table 4 and Figure 5).

This result does not support hypothesis 1 that values influence employee satisfaction in a precise manner. However, as noted above, characteristic values differ by ES level and employment pattern. Thus, the result implies that values influence ES.

4.3. Dependency Word Map Analysis. A dependency word map analysis was then performed based on shop, staff/employee, and customer, ranking the top three nouns in terms of appearance. These contexts in keywords were used in different ways (Figures 6, 7, and 8). The context in this

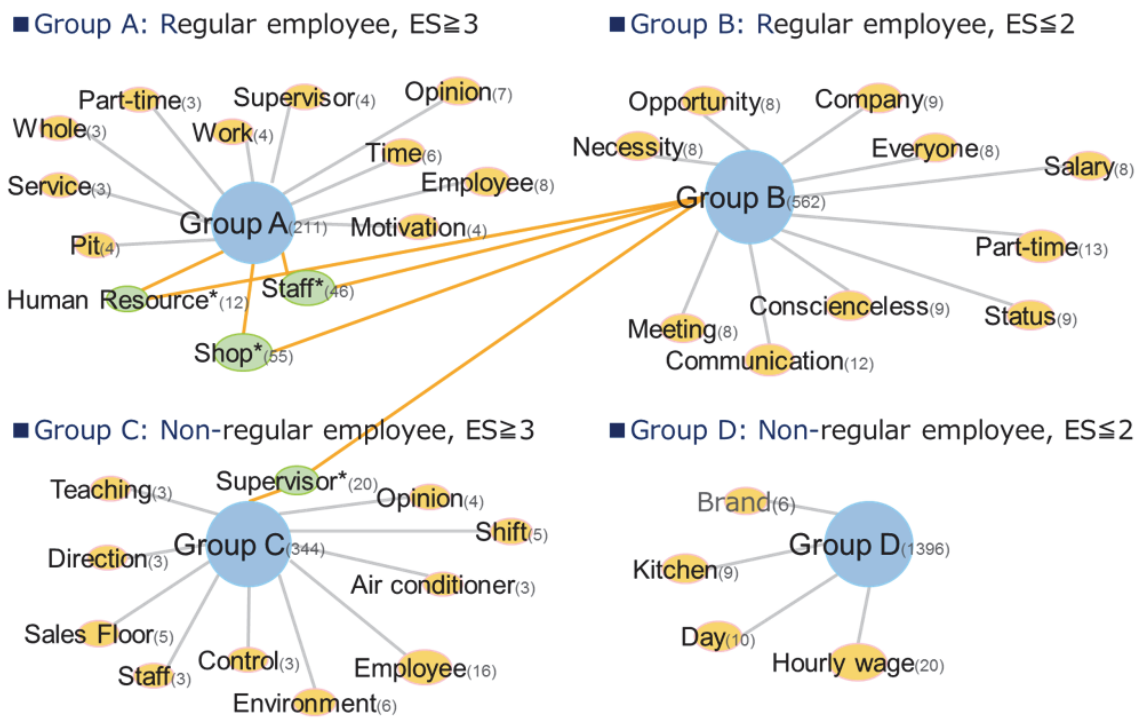


Figure 5: Keyword map by employee satisfaction/employment pattern.

context is to understand the values in the background from the connection between words and words.

The blue circles indicate the top three nouns. The green circles with gray lines display words that depend on the top three words. A gray arrow represents a relationship. Beginning at each word, arrows connect the most relevant words employed in conjunction with the former word. The numbers indicate the number of times a word was found in the comments.

4.4. Characteristic Items by Employment Pattern. Extraction of common keywords used by regular and nonregular employees yielded the data in Tables 5 and 6. The characteristic keywords showed different patterns. The total ranking of keywords are shown in Table 7.

Regular employees often commented on the organization, while nonregular employees tended to focus on their own personal problems. Employees with a high degree of satisfaction frequently commented on people, whereas low-satisfaction employees tended to list physical things such as salary.

5. Discussion

This result of correlation coefficients in Section 4.1 shows that the company motivates regular and nonregular workers by the same way to gain eternal ES factors despite the different positions. However, some nonregular workers may desire salaries rather than evaluation and growth. That is why it is important to focus on internal ES factors and analyze unique factors of regular and nonregular employees to really motivate them.

Text mining revealed differences between employees with a new perspective. Section 4.2 shows that the greater the employee satisfaction, the greater the interest in personal improvement and work. In the retail and service industries, it is extremely important to communicate with colleagues daily, often on a site-by-site basis. The higher the level of interest, the more challenging the task. However, retail and service shopfront operations are monotonous and stressful if employees lack interest in their future. Thus, hypothesis 1 (values influence employee satisfaction) was supported. The occurrence rates of 93 keywords cited by high- and low-ES employees were compared using the t-test. The rates were statistically significant for the nonregular employees ($P < 0.001$, two-sided).

Table 7 showed that even common words such as "shop", "employee", and "customer" assumed different meanings depending on employment pattern. Regular employees adopted the view of the organization and considered relationships in terms of "level" and "communication", whereas nonregular employees linked the words to "environment" and "toilet", thus viewing customers as individuals. Thus, the values of regular and nonregular workers differed. Table 7 showed that the principal keywords of regular and nonregular employees also differed. Regular employees suggested management improvements; they wished to be considered as more than "human resources" or the "company"; keywords used by nonregular employees often related to daily work. Thus, regular and nonregular employees differed in terms of work and workplace values, and in how to improve the workplace, supporting hypothesis 2. Thus, different labor orientations are in play even in the absence of any correlation between employee satisfaction and various factors.

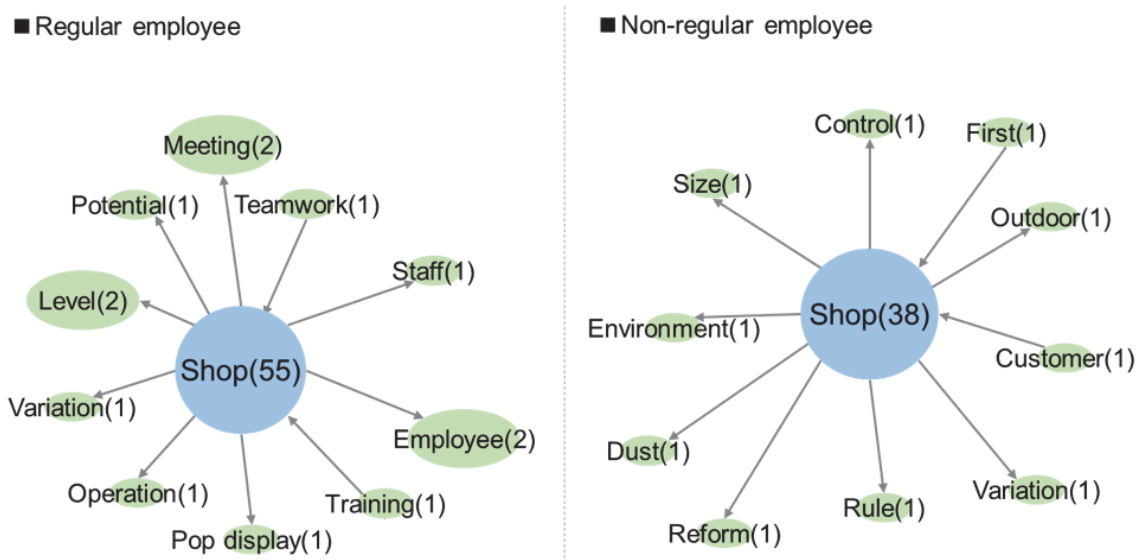


Figure 6: **Dependency word map of nouns by employment pattern regarding "Shop"**. Regular employees recognized improvement requirements in their organizations. For example, improvements in meetings, teamwork, and employees. Nonregular employees often commented on customers and the daily environment, including such words as customer, rule, and environment.

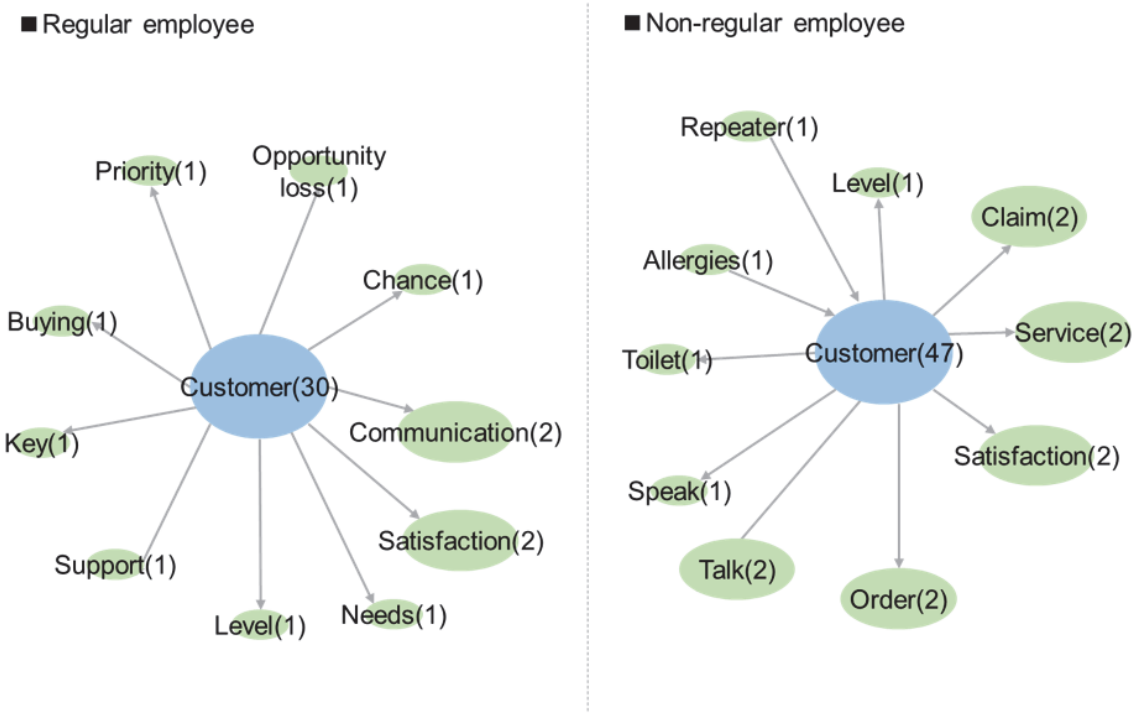


Figure 7: **Dependency word map of nouns by employment pattern regarding "Customers."** Regular employees often used keywords associated with the future, such as chance, opportunity, or priority. In contrast, nonregular employees included many keywords for the present, such as order, service, or claim.

Table 5: Regular employees (n = 270).

Keyword	n	Rate	Representative comment (underlined words indicate keywords.)
Human resources	19	7.0%	There are few human resources and many are not hospitable. There seems to be a lack of customers who are highly motivated to purchase.
Company	12	4.4%	Company management is concerned only with profit. For that reason, employees' wages are low and motivation is falling.
Consciousness	10	3.7%	I feel that consciousness of one's position as a professional salesperson can be connected to customer satisfaction by greater sharing of promotional ideas.
Individual	7	2.6%	We should develop individual skills, respect human nature, and use as many coaching strategies as the number of employees.
Headquarters	6	2.2%	I said at the meeting with my boss at headquarters that I would improve; I will not hesitate. I understand that there are too few people. In case...
Field	5	1.9%	Please understand why many people leave and listen to the voices of employees in the field.

Many keywords pertain to the entire organization rather than daily life in the workplace.

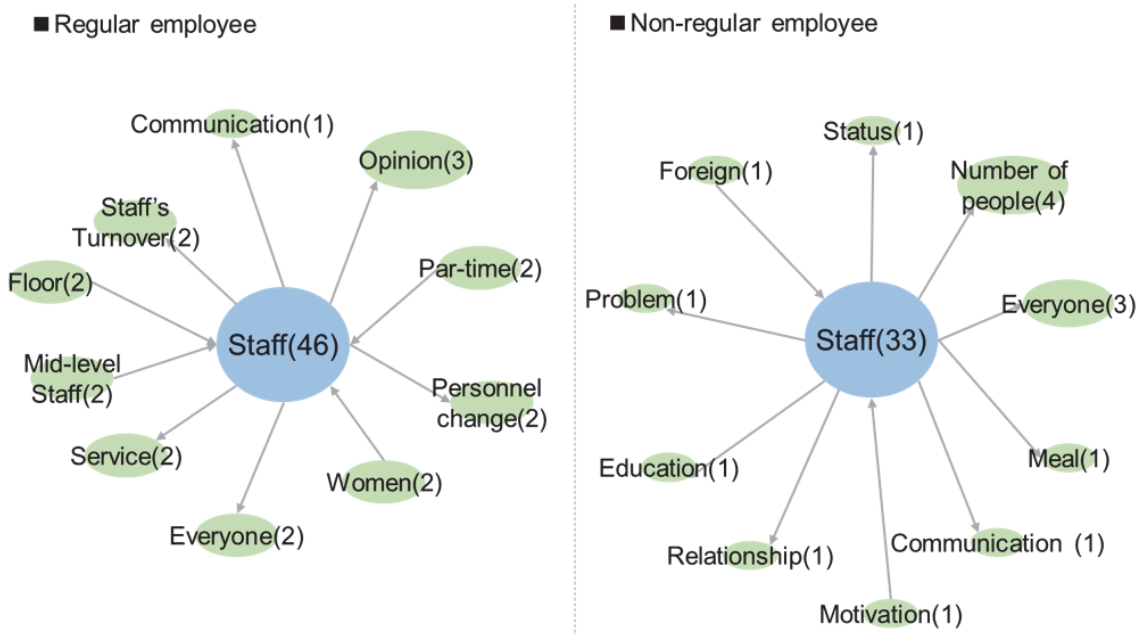


Figure 8: **Dependency word map of nouns by employment pattern regarding "Staff".** Regular employees used keywords concerning human resources measures related to the organization. These keywords included staff turnover, mid-level, and personnel change. Nonregular employees used keywords that related to their own issues surrounding their daily work.

The occurrence rates of 93 keywords cited by regular and nonregular employees were also compared using the t-test. No statistically significant difference was found for the nonregular employees (P = 0.128, two-sided).

Recently, the MHLW [25] released "Draft guidelines for same-labor, same-wage" seeking to improve the treatment of nonregular employees; employment should be regular. However, the focus here is not on improvement of working conditions but, rather, on motivation patterns. Companies can match human resources to workplace needs by recognizing employees' work. It is also necessary to convey companies' philosophies to the employees. If they are able to work in a workplace that offers a good fit for their values, many nonregular employees will not leave. This is more powerful than support from extrinsic factors.

6. Conclusion

In this paper authors explore the effects of the values and attitudes of retail and service industry employees on ES and identified differences between regular and nonregular employees. Employee values affected ES; the values of regular and nonregular employees are not significantly statistically different. However, keywords of free-answer comments implied the values of both features.

Table 6: Nonregular employees (n = 383).

Keyword	N	Rate	Representative comment
Hourly wage	22	5.7%	Everyone is suffering from low hourly wages and the newsletter says that some people worry; I am one of them.
Goods	19	5.0%	The goods are of poor quality. If items that are dirty or broken are replaced or made more attractive, I think that the quality of the shop would go up.
Kitchen/hall	19	5.0%	I often feel the communication gap between the hall and the kitchen, so I want to communicate more and improve the atmosphere of the shop.
Newcomer	12	3.1%	There are almost no manuals at this store. I think that everyone, privately, agrees with this. I think there are some good aspects here, but for newcomers, there may also be a lot of confusion.
Shifts	10	2.6%	I think that you should be strict, and not careless about time; insist that shifts are kept, so that the physical condition of the store can be managed.
Toilet	7	1.8%	I think that a toilet check would be a good idea for the business. The toilet is always in a miserable state.
Members	7	1.8%	I think the differences in the skills of members should be improved.
Order	5	1.3%	I do not mind taking over work or giving orders, but I feel that this has become a vicious circle recently.

Many keywords expressed views based on the daily work at the site rather than that of the whole organization.

Table 7: Keywords in free comments on proposals for improving stores (top 20 nouns).

Regular employee	N	%	Non-regular employee	N	%
Shop＊＊＊	71	26.3	Staff/employee＊＊＊	86	22.3
Staff/employee＊＊＊	70	25.9	Shop＊＊＊	69	17.9
Customer＊＊＊	46	17.0	Customer＊＊＊	47	12.2
Manager	30	11.1	Non-regular employment	41	10.6
Work	24	8.9	Work	30	7.8
Non-regular employment	17	6.3	Salary	22	5.7
Communication	16	5.9	Products＊＊	19	4.9
Opinion	15	5.6	Manager	25	6.5
Time	15	5.6	Communication	15	3.9
Company＊	12	4.4	Environment＊＊	12	2.0
Human resource＊	12	4.4	Education	12	2.0
Motivation＊	10	3.7	Time	12	2.0
Consciousness＊	10	3.7	Newcomer＊＊	12	2.0
Improvement＊	10	3.7	Opinion	11	1.8
Salary	10	3.7	Salary	11	1.8
Opportunity	9	3.3	Way＊＊	11	1.8
Education	9	3.3	Kitchen＊＊	10	1.7
Status	9	3.3	Shift＊＊	10	1.7
Individual	9	3.3	Place	10	1.7
Meeting	8	3.0	Days	10	1.7
Environment	8	3.0			

Question: Please tell how you would make your store better via innovations and improvements. Percentages of all respondents who used the words to the left. Respondents: regular employees (n=270); nonregular employees (n=383). In the table above, the number of uses at 20th was the same, so 21 keywords are posted.
＊Keywords frequent only in regular employee comments.
＊＊Keywords frequent only in nonregular employee comments.
＊＊＊Keywords frequent (over 10%) in both regular and nonregular employee comments.

The findings of this study suggest a management model of service industries. However, the model should not be always applicable to all service industries. Combinations of text mining and quantitative methods will yield data that are more accurate; such work is planned.

Data Availability

The.xlsx data used to support the findings of this study are available from the corresponding author upon request. However, since some text information in the original data includes personal information, some text may be edited before being provided.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are deeply grateful to the companies and store employees who responded to the study questionnaire. The authors would also like to thank the employees of Recruit Management Solutions, who provided helpful comments and suggestions related to new developments in survey techniques. The authors also thank the Foundation for the Fusion of Science and Technology for a research grant that made this study possible.

Supplementary Materials

Supplementary Materials shows trends in the labor force population ratios around the world. And the ratio generally refers to the population of 15 to 65 years old or 15 to 60 years old) in the total population in each country. This data supports the fact that the decline in the labor force in Japan is remarkable compared to other countries. *(Supplementary Materials)*

References

[1] Ministry of Health, Labour and Welfare. *Vision of working habits*, 2012.

[2] The Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, *World Population Prospects*, 2015.

[3] Japan Productivity Center, *New employee "New employee" Consciousness of work Survey result*, 2012.

[4] A. R. Hochschild, "Emotion Work, Feeling Rules, and Social Structure," *American Journal of Sociology*, vol. 85, no. 3, pp. 551–575, 1979.

[5] Ministry of Health, Labour and Welfare. *Study Group for the actual conditions on the difference or the like of the treatment by the employment of worker*, 2017.

[6] Ministry of Health, Labour and Welfare. *Analysis of Labor Economy* 2017.

[7] The Japan Institute for Labor Policy and Training, *Non-regular Employment-issues and Challenges Coon to the Major Development Countries*, 2011.

[8] C. E. Crangle, "Text Summarization in Data Mining," in *Soft-Ware 2002: Computing in an Imperfect World*, vol. 2311 of *Lecture Notes in Computer Science*, pp. 332–347, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[9] C. P. Alderfer, "An empirical test of a new theory of human needs," *Organizational Behavior and Human Decision Processes*, vol. 4, no. 2, pp. 142–175, 1969.

[10] F. Herzberg, "One More Time: How Do You Motivate Employees?" *Harvard Business Review*, vol. 81, no. 1, pp. 87–141, 2003.

[11] J. K. Harter, F. L. Schmidt, and T. L. Hayes, "Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis," *Journal of Applied Psychology*, vol. 87, no. 2, pp. 268–279, 2002.

[12] D. Carter and T. Baghurst, "The Influence of Servant Leadership on Restaurant Employee Engagement," *Journal of Business Ethics*, vol. 124, no. 3, pp. 453–464, 2014.

[13] J. P. Eggers and A. Kaul, "Motivation and ability? A behavioral perspective on the pursuit of radical invention in multi-technology incumbents," *Academy of Management Journal (AMJ)*, vol. 61, no. 1, pp. 67–93, 2018.

[14] K. Pugliesi, "The consequences of emotional labor: Effects on work stress, job satisfaction, and well-being," *Motivation and Emotion*, vol. 23, no. 2, pp. 125–152, 1999.

[15] S. H. Schwartz, "Are there universal aspects in the structure and contents of human values?" *Journal of Social Issues*, vol. 50, no. 4, pp. 19–45, 1994.

[16] M. Ros, S. H. Schwartz, and S. Surkiss, "Basic individual values, work values, and the meaning of work," *Applied Psychology*, vol. 48, no. 1, pp. 49–71, 1999.

[17] J. M. Twenge, "A review of the empirical evidence on generational differences in work attitudes," *Journal of Business and Psychology*, vol. 25, no. 2, pp. 201–210, 2010.

[18] S. M. To and H. L. Tam, "Generational Differences in Work Values, Perceived Job Rewards, and Job Satisfaction of Chinese Female Migrant Workers: Implications for Social Policy and Social Services," *Social Indicators Research*, vol. 118, no. 3, pp. 1315–1332, 2014.

[19] L.-Y. Sun, S. Aryee, and K. S. Law, "High-performance human resource practices, citizenship behavior, and organizational performance: a relational perspective," *Academy of Management Journal (AMJ)*, vol. 50, no. 3, pp. 558–577, 2007.

[20] J. L. Heskett, T. O. Jones, G. W. Loveman et al., "Putting the Service-Profit Chain to Work," *Harvard Business Review*, 1994.

[21] T. M. Amabile and S. J. Kramer, "Inner work life: Understanding the subtext of business performance," *Harvard Business Review*, vol. 85, no. 5, pp. 72–83, 2007.

[22] E. W. Morrison, "Employee Voice and Silence," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 1, no. 1, pp. 173–197, 2014.

[23] T. Matsuki and J. Nakamura, "Effects of employee satisfaction on service by employment pattern," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017.

[24] J. Ford, D. Nierle, P. Leeds, and T. Stetz, "Text Mining Narrative Survey Responses to Develop Engagement Scale Items," in *Proceedings of the 51th Hawaii International Conference on System Sciences*, 2018.

[25] Ministry of Health, Labour and Welfare. *Proposal of same labor same wage guideline*, 2016.

Hindawi
Advances in Human-Computer Interaction
Volume 2018, Article ID 9054685, 12 pages
https://doi.org/10.1155/2018/9054685

Research Article

How to Understand Belief Drift? Externalization of Variables Considering Different Background Knowledge

Teruaki Hayashi and Yukio Ohsawa

Department of Systems Innovation, School of Engineering, Tokyo, Japan

Correspondence should be addressed to Teruaki Hayashi; teru-h.884@nifty.com

Received 28 June 2018; Revised 1 November 2018; Accepted 21 November 2018; Published 4 December 2018

Guest Editor: Rafal Rzepka

It is necessary to make decisions by integrating appropriate information that is not used in daily life in disaster prevention before, during, and after disasters. Despite this, it is difficult for people to make use of appropriate information under circumstances where various kinds of information are complicated. People can be in an agitated state in which they do not know what will happen. In this paper, we define this situation as Belief Drift (BD) and discuss what kinds of data should be acquired to understand situations of BD because factors causing BD may be diverse. We collected explanations of BD from researchers with different background knowledge and discussed sets of variables inferred by VARIABLE QUEST (VQ). VQ is the inferring method for variables unifying cooccurrence graphs of variables in the datasets. The results indicate that common variables are externalized from the different explanations of BD by researchers with different background knowledge. Results suggest that, even if the terms used to explain the state of BD differ, the data acquired to understand BD are common.

1. Introduction

Projects seeking to protect society from disasters are progressing globally. The Department of Homeland Security, the Department of Energy, and so on in the United States have conducted infrastructure protection projects since 2003 [1, 2]. Concerning predicted disasters, technologies to evaluate risks have been developed systematically, including vulnerabilities and interdependencies of 12 essential infrastructures. Conversely, the Japanese government has tackled problems of information infrastructure. Resilient systems considering a variety of cellular phones and smartphones have been studied academically and practically from the viewpoints of the possibility of message transmission, stability, and the reliability of information [3, 4]. Especially for disaster prevention ahead of the Tokyo 2020 Olympic and Paralympic Games, the government has provided the Disaster Prevention Portal for foreigners [5].

However, while the robustness and resilience of the infrastructure have increased, victims of natural disasters have not been able to reach favorable decisions because of a state of desperation from anxiety. The kinds of information that have been delivered to those who were anxious during and after disasters constitute an urgent issue. For example, in the examinations of internal radiation exposure after the accident of the First Nuclear Power Plant in the Fukushima Prefecture, the affected area was divided into residents that were hardly affected by radiation exposure and those that were noticeably affected [6, 7]. In the medical examinations after the accident, rapid chronic diseases were evident [8]. Nara mentioned that, although recognition and anxiety exist, the information transmitted by risk management institutions cannot gain the confidence of Japanese residents [9]. Information and the media are diverse, which makes it difficult to obtain consistent messages [10]. Even in the Chernobyl nuclear power plant accident, the relationship between the accident and carcinogenesis is scientifically uncertain [11].

In disaster prevention before, during, and after disasters, it is crucial to create a methodology that provides appropriate integrated information used in daily life. In such situations, people cannot establish certain beliefs because of inconsistent and diverse information. For example, when inconsistent information is given such as "that town is polluted and cannot live," or "this city is safe" from disasters, it evokes

anxiety and anger in information senders and surrounding people and causes more meaningless messages to proliferate [12]. Research mentions that anxiety amplifies continuously and has taken root among residents. Depending on conditions, differing information may spread as inconsistent information, which confuses people because they cannot be confident as to what to believe. We defined this situation as Belief Drift (BD) and began this project to study this [13]. The essential mission of our project is to establish the fundamental methodology and systems of information generation, propagation, acceptance with reliability, usefulness, and consistency for people who are anxious during and after disasters.

To detect BD and create a system providing appropriate information for people whose beliefs are drifting, we acquired data to understand situations of BD. However, there is a problem of a vocabulary gap among different researchers. When disciplinary fields differ, there are many different factors to consider in actions. Based on these challenges, because BD is a multidisciplinary question, it is difficult to form a common recognition if background knowledge of researchers differs. Also, if background knowledge differs, the representation of situations also differs, which makes it difficult to decide what kind of data acquisition is sufficient to achieve a purpose.

In this paper, as a precursor for creating our methodology, we collected explanations of BD from researchers with different background knowledge and discussed what kind of data (sets of variables) we should acquire to detect BD using VARIABLE QUEST (VQ). VQ is the method of inferring Variable Labels unifying cooccurrence graphs of variables in the datasets. Variable Labels (VLs) are the names/meanings of variables in datasets. Our approach is to input the explanations of BD to VQ and obtain sets of VLs as data related to BD. We compared and discussed the differences among terms in the explanations and obtained VLs to understand the different perspectives of researchers.

The remainder of this paper is organized as follows. In Section 2, we explain the details of our approach to detect the situation of BD. In Sections 3 and 4, we show the techniques used in the experiment, i.e., Data Jackets (DJs) and VQ. Section 5 describes the purpose and experimental details. Section 6 shows the preliminary experiment for testing the performance and the reproductivity of our proposed approach using test data. In Section 7, we show the results of the analysis, and we discussed them in Section 8. Finally, Section 9 concludes with a brief review and discussion of future work.

2. Our Approach

Humans have acquired and used various data for making decisions in business, politics, or, even, daily activities. However, agents in different fields have different background knowledge. They sometimes use different terminology to explain the same concept or event. Conversely, different agents use the same term to understand different events globally. Thus, it is possible that recognition of the situations may differ by agents' background knowledge. From the studies in cognitive science, it was shown that two problem solvers might construct different facts even if they observe the same data because of the different perspectives provided by their contexts and background knowledge [16]. Metcalfe explained the same phenomenon from the field of economics. Decision makers in society make the most rational choices individually. However, they may recognize different worlds, even though they see the same world because of their background knowledge and available opportunities [17]. Boisot and Canals explained the difference between data, information, and knowledge, and specific types of utility [14]. They demonstrated that data is a property of events and things in the world, and information, by contrast, depends on expectations or states of knowledge.

Figure 1 represents the different conditions of data, information, and knowledge of the agent globally. The model was proposed to understand data, information, and knowledge as different economic factors. Based on this model, the recognition of the agent globally proceeds as follows.

Step 1. World events produce stimuli.

Step 2. The agent receives external stimuli through perceptual filters and acquires data (note that perceptual filters have limitations and cannot attain all stimuli from events).

Step 3. Obtained data are converted into information through conceptual filters (considering the previous studies, mechanisms by which different agents may recognize different worlds while looking at the same event are affected by the conceptual filters).

Step 4. The agent refers to background knowledge and recognizes events globally (the actions toward the world in Figure 1). Also, the recognition gives feedback on the perceptual and conceptual filters (the expectations from the agent's knowledge to each filter in Figure 1).

The purpose of this study is to understand BD. The kinds of data we should observe and collect to understand BD are problematic. It is essential to understand what kind of data the agent acquires to understand unknown phenomena. In this study, the agent is an observer, and the unknown phenomena are the BD. When we apply this model to our research subject, there is a possibility that the data acquired for understanding BD with different background knowledge may differ. Conversely, there is a possibility that the data considered essential for observing BD is common even if agents have different background knowledge. Based on the above discussion, we can summarize the hypotheses of this paper as follows:

Hypothesis 1. Even if the agents have different background knowledge in understanding BD, important data for understanding BD is common.

Hypothesis 2. Because the agents have different background knowledge to understand BD, important data for understanding BD differs.

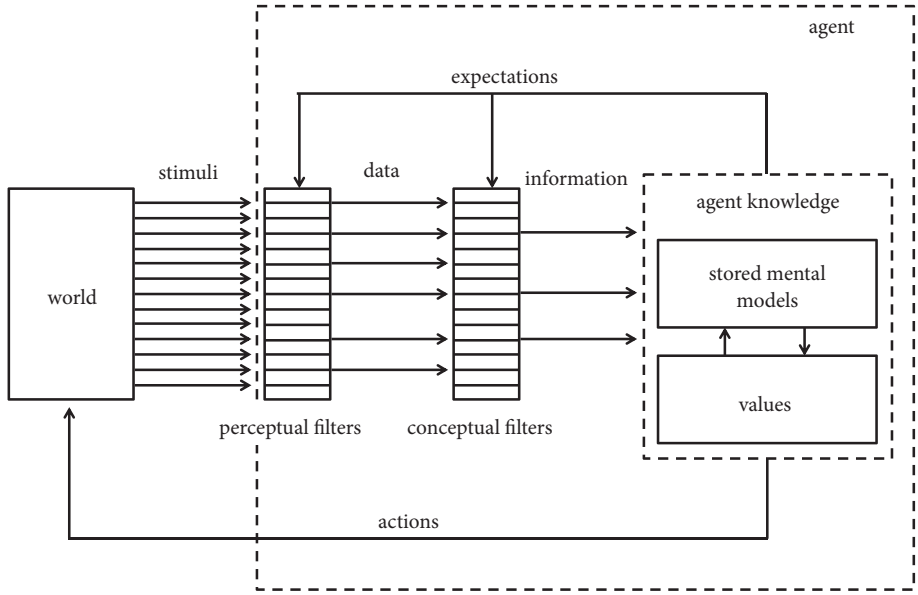


Figure 1: The agent-in-the-world model [14].

To verify the hypotheses above, we collected explanations of BD from researchers with different background knowledge and compared the terms used to explain BD. Because it is difficult to observe the background knowledge of the researchers directly, we assume that explanations of BD represent background knowledge. Also, it is difficult to observe information on data important for understanding the events directly; we use Data Jackets (DJs) as summary information on data and Variable Labels (VLs) in DJs as the detailed information about variables in data. The detailed explanations about DJs and VLs are explained in the next section.

3. Data Jacket (DJ)

Data Jacket (DJ) is a technique used for sharing information about data and for considering the potential value of datasets with the data being hidden. The idea of DJ is to share "a summary of data" in natural language as meta-data without sharing specific data [18]. Sharing the summaries of data as DJs enables data holders to provide information on their data, reducing the risk of data management, cost, and privacy. Also, data users can easily find data related to their interests through descriptions of DJs [19]. Table 1 is an example of DJ of "Vegetable Production in Japan." In DJs, variables are described by Variable Labels (VLs). A VL is the name/meaning of variables in datasets. In DJs, variables in data are summarized as VLs, which are the meta-data of variables and values in datasets. For example, the dataset "Vegetable Production in Japan" includes VLs "location of producer," "weekly (or even daily) production of each producer," "weekly expenditure on vegetable production," "selling prices," and "the type of vegetable."

In this research, background knowledge is given by sentences (set of terms), and the data important to observe the situation of BD requires a set of variables. DJs and VLs are summarized information described in natural language, and over 1,000 pieces of information on data and roughly 5,000 VLs have been collected from different domains. Although larger published databases such as DBpedia [20] are provided as Linked Open Data, they specialize in publicly available data. DJ is not limited to public data and contains information about variables from private companies and individuals. To understand the situation of BD, it is reasonable to use the dataset including various information on variables.

In this paper, to understand what kinds of data (sets of VLs) should be acquired to understand situations of BD, we use the outlines of data and VLs in DJs as corpus data of VARIABLE QUEST explained in detail in the next section.

4. VARIABLE QUEST (VQ)

4.1. The Overview of VQ. VARIABLE QUEST (VQ) is the network visualization system of VLs using the matrix-based inferring method of VLs [15, 21]. VQ represents the cooccurrence and the frequency between VLs in DJs. The cooccurrence of VLs is a feature in which there is a highly frequent pair of VLs appearing simultaneously in the data, e.g., "latitude" and "longitude," or "name," "age," and "nationality." VQ introduces the function of the fundamental matrix-based algorithm to infer VLs from outlines of data (ODs) whose VLs are missing or unknown. VQ has two important models to infer VLs as follows.

Model 1. Datasets are similar when their information for explaining data is similar.

Model 2. Datasets have similar VLs when the similarity of datasets are higher.

Table 1: An example of DJ and VLs.

Item	Content
Title of data	Vegetable Production in Japan
Outline of data (OD)	Vegetables are expensive in Japan compared with other countries. Therefore, the production of vegetables is significant for this country. This database records detailed information on the location, quantity, type, and time of vegetable production. Certain institutions could use the data to analyze factors that impact vegetable price and production. Methods could be created to reduce prices, or to profit from reasonably allocating production.
Variable Labels (VLs)	"Location of producer," "Selling prices," "The type of vegetable," "Weekly expenditure on vegetable production," "Weekly (or even daily) production of each producer."
Sharing Policy	Undecided
Formats of data	CSV
Types of data	"Numerical values," "Table," "Text"
How to collect data	Collect from individual sellers in the local market, agricultural firms, and distributors.

Table 2: The examples of top-ten VLs inferred by VQ.

OD_1		OD_2	
Inferred VLs	Similarity	Inferred VLs	Similarity
Number of births	0.349	North latitude (degrees)	0.361
Number of deaths	0.349	Installer of the device	0.361
In-migrants	0.335	Address of the seismic intensity	0.361
Fatalities	0.335	East longitude (degrees)	0.361
Out-migrants	0.335	North latitude (minutes)	0.361
Population	0.321	The name of the seismic intensity	0.361
Number of households	0.318	The pronunciation of seismic intensity	0.361
Population (male)	0.318	Match level	0.361
Population (female)	0.318	Earthquake number	0.360
Fertilities	0.318	Position number	0.360

VQ introduced the bag-of-words and vector space model [22] for creating the corpus from the training data (DJs with VLs). In the preprocessing steps, VQ conducts the morphological analysis of the text of ODs, extracting words, removing stop words, and restoring words to their original forms.

Table 2 shows two examples of inferred VLs. The left column of Table 2 shows the top-ten inferred list of VLs obtained from an OD_1; "this data represents the transition of the population of each year in Tokyo, Japan." The right column is the VLs found from an OD_2 "the earthquake data in the world." We can obtain a set of VLs with similarities to the queries whose VLs are unknown. Even if a free text query does not include terms which represent VLs, VQ returns related sets of VLs with the query.

4.2. Detailed Algorithm of VQ. In this subsection, we explain the detailed algorithm of VQ. At first, VQ conducts an algorithm to calculate the similarity among training data of ODs based on Model 1. ODs are given by the sentences so that we assume that each OD is a set of terms. After conducting the preprocessing steps to ODs using a bag-of-words model, the ODs are converted into a matrix representation (a Term-OD matrix). A Term-OD matrix M ($W \times D$) consists of W-dimensional OD vectors as columns and D-dimensional term vectors as rows. Each element in the matrix M (v_{ij}) in an OD vector ($\boldsymbol{od}_j$) corresponds to the frequency with which a term (a row i) occurs in an OD (a column j) as shown in (1) and (2). Note that the subscript T on the upper-right corner of vectors represents the transposition, and the vectors are highlighted in bold in this paper.

$$M = \left(\boldsymbol{od}_1, \cdots, \boldsymbol{od}_j, \cdots, \boldsymbol{od}_D\right) \tag{1}$$

$$\boldsymbol{od}_j = \left(v_{1j} \ \cdots \ v_{ij} \ \cdots \ v_{Wj}\right)^{\mathrm{T}} \tag{2}$$

In the second step, a set of VLs in DJs is converted into a VL-OD matrix R ($V \times D$). In the training data of DJs, ODs and VLs are linked when they appear in the same DJs. Each element in the matrix R (r_{ij}) in the jth OD vector ($\boldsymbol{od'}_j$) corresponds to the frequency (0 or 1) with which the ith VL occurs in the jth OD as shown in (3) and (4).

$$R = \left(\boldsymbol{od'}_1, \cdots, \boldsymbol{od'}_j, \cdots, \boldsymbol{od'}_D\right) \tag{3}$$

$$\boldsymbol{od'}_j = \left(r_{1j} \ \cdots \ r_{ij} \ \cdots \ r_{Vj}\right)^{\mathrm{T}} \tag{4}$$

In the third step, a Term-VL matrix E ($= MR^{\mathrm{T}}$) is generated by combining a Term-OD matrix M and an OD-VL matrix R^{T}. The Term-VL matrix E is represented by (5), and the jth

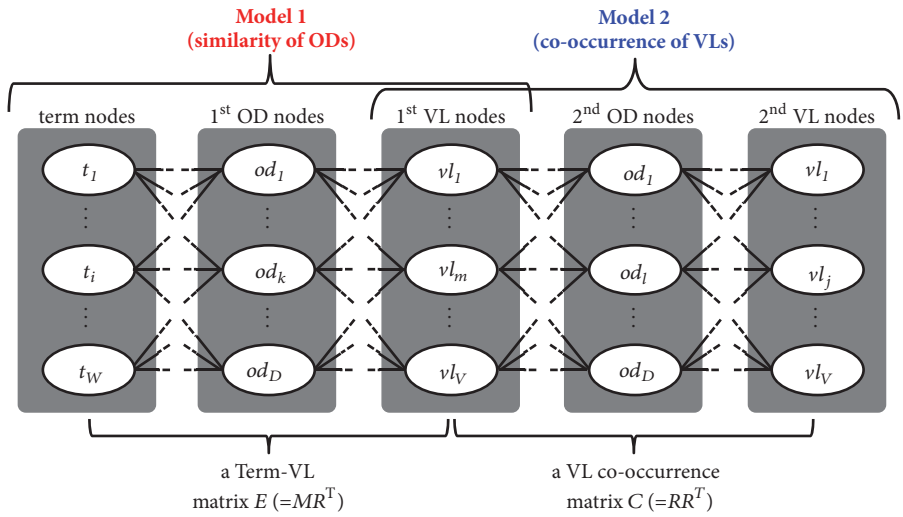


FIGURE 2: The structure of the Term-VL matrix *EC* [15] (partially modified by authors).

VL vector ($\boldsymbol{vl}_j$) is given by (6). The elements of the Term-VL matrix E (e_{ij}) are calculated by (7).

$$E = \left(\boldsymbol{vl}_1, \cdots, \boldsymbol{vl}_j, \cdots, \boldsymbol{vl}_V\right) \tag{5}$$

$$\boldsymbol{vl}_j = \left(e_{1j} \ \cdots \ e_{ij} \ \cdots \ e_{Wj}\right)^{\mathrm{T}} \tag{6}$$

$$e_{ij} = \sum_{k=1}^{D} v_{ik} r_{kj} \tag{7}$$

Each element in the Term-VL matrix E means the sum of the product of the frequency (v_{ik}) with which the ith term (t_i) occurs in the kth OD (od_k) and the frequency (r_{jk}) with which the jth VL (vl_j) links with the kth OD (od_k).

In the fourth step, we conduct the VL cooccurrence matrix C (= RR^{T}), assuming that any pair of VLs in the same DJ occurs once based on Model 2. The elements in the VL cooccurrence matrix C (c_{ij}) represent the number of DJs which include a pair of the ith VL (vl_i) and the jth VL (vl_j) as shown in (8).

$$c_{ij} = \sum_{k=1}^{D} r_{ik} r_{kj} \tag{8}$$

In the fifth step, we acquired a Term-VL matrix *EC* by a product of the Term-VL matrix E and the VL cooccurrence matrix C. The Term-VL matrix *EC* consists of V-dimensional term vectors as rows and W-dimensional VL vectors as columns. The structure of the Term-VL matrix *EC* is the same as that of the Term-VL matrix E. The element g_{ij} of the matrix *EC* is given by (9). The value is calculated by the similarities of ODs and queries, which is the function of the matrix E, and the cooccurrence of VLs, which is the function of the matrix C.

$$g_{ij} = \sum_{m=1}^{V} \left(\sum_{k=1}^{D} v_{ik} r_{km}\right)\left(\sum_{l=1}^{D} r_{ml} r_{lj}\right) \tag{9}$$

The structure of the Term-VL matrix *EC* is equivalent to the adjacency matrix of the 5-partite graph as shown in Figure 2. When OD_x whose VLs are unknown is inputted, VQ calculates a W-dimensional feature vector of OD_x ($\boldsymbol{od}_x$) referring to the dictionary (the list of terms) and the corpus. By calculating the similarities of feature vector of OD_x ($\boldsymbol{od}_x$) and each feature vector of VL ($\boldsymbol{vl}_j$) by the matrix *EC*, VQ returns a scored set of VLs with similarities. In this paper, the similarity scores of $\boldsymbol{od}_x$ and $\boldsymbol{vl}_j$ are calculated by cosine similarities (similarity($\boldsymbol{od}_x, \boldsymbol{vl}_j$) = $\boldsymbol{od}_x \cdot \boldsymbol{vl}_j / |\boldsymbol{od}_x||\boldsymbol{vl}_j|$).

In this paper, we use VQ to externalize VLs related to the state of BD, even if the terms explaining BD differ among researchers. In the next section, we explain the purpose and the method of our experiment.

5. Experimental Details

5.1. Purpose and Method. The purpose of this experiment is to acquire sets of VLs necessary to understand the situation of the users whose beliefs have drifted. To achieve this goal, we obtained linguistic sentences to explain the state of BD from researchers with different background knowledge. Using VQ with VLs, we acquired relevant data (a set of VLs) from sentences concerning BD. We attained linguistic sentences from six researchers to explain the state of BD. Three researchers are from the department of engineering, two are medical doctors, and one is a psychotherapist. All researchers hold Ph.D. in their fields. Researchers have discussed BD through several meetings for roughly one and a half years. Although six samples are relatively small, because BD is a new research topic, such samples cannot be collected elsewhere. To supplement the small number of samples, we introduce the index of commonality (10) instead of quantitative evaluation and compared degrees of variation.

We asked the researchers to write down "the state of Belief Drift you understand from your own background knowledge." To avoid the bias of others' answers, we asked them to submit the sentences separately. We input these

Table 3: Corpus statistics of VQ (parentheses represent the standard deviation).

Number of DJs	1,032
Total number of terms in DJs	27,194
Unique terms in DJs	4,971
Mean of the number of terms in each DJ	26.4 (58.5)
Maximum number of terms in DJs	1471
Minimum number of terms in DJs	1
Total number of VLs in DJs	7,029
Unique VLs in DJs	5,155
Mean of the number of VLs in each DJ	6.81 (8.80)
Maximum number of VLs in DJs	118
Minimum number of VLs in DJs	1

answers to VQ as queries described in Section 4 and obtained sets of top-30 likely VLs. Note that explanations of BD and some VLs are given in Japanese. In this paper, we translated all terms into English after the experiment.

5.2. Datasets and Method for Evaluations. To construct the corpus for VQ, we use 1,032 DJs including outlines of data (ODs) and VLs, collected from business persons, researchers, and data holders. All DJs are extracted from DJ Store which is a database with a retrieval system for DJs on the Web and is provided in RDF format [23]. Table 3 shows the statistics of corpus data. Each DJ has 6.8 VLs on average. 5,155 unique VLs are stored in total. The corpus and the dictionary were constructed from all the words in OD texts. The OD corpus consists of 4,971 unique words. We used MeCab for the morphological analysis [24], which is a common tool for analyzing morphemes of Japanese texts.

For weighing discriminative terms in the corpus, we used tf-idf in a weighting scheme [25], which is reliable for identifying distinctive terms in documents. The term frequency (tf) is the number of times a term appears in a document, and the inverse document frequency (idf) diminishes the weight of frequent terms in all documents and increases the weight of those terms which appear rarely. When inputting the sentences about BD in VQ, we removed punctuation marks and symbols in the texts as stop words, restored words to their original forms, and extracted nouns, verbs, adverbs, and adjectives as a preprocess.

To evaluate the commonality among researchers, we define the indicator of commonality. Equation (10) is an indicator for evaluating the degree of commonality of elements among clusters, which calculates the proportion of excluding terms appearing only once. T_i represents the ith set of elements, and $|T_i|$ represents the number of elements in the ith set T_i. Note that $n = 6$ and *ele* is term.

$$\text{commonality}(ele) = 1 - \frac{\sum_{i=1}^{n} \left| T_i - \bigcup_{j=1, j\neq i}^{n} T_j \right|}{\left| \bigcup_{i=1}^{n} T_i \right|} \quad (10)$$

To test the performance and the reproductivity of our approach, we first conduct the preliminary experiment in the next section by using the test data.

Table 4: Commonalities of terms and VLs using test data at random (parentheses represent the standard deviation).

Commonality	Terms	VLs
Mean	0.116 (0.037)	0.055 (0.041)
Median	0.112	0.045
Maximum value	0.208	0.200
Minimum value	0.037	0.000

6. Preliminary Study

6.1. Experiment. At first, by using the same corpus of DJs shown in Table 3, we compared the commonality of terms in DJs and VLs obtained from VQ. The purpose of this study is to understand the kinds of data we should observe and collect to understand the situation of BD. To adjust to this goal, we set three different themes ("transportation," "health," and "personal activity"). We extracted 7 DJs related to each theme, obtained top-30 VLs, and calculated the commonalities of terms and VLs. To compare the performance, we also calculated the commonalities of terms and VLs randomly. We repeated the following steps for 80 times: (1) choosing 10 DJs at random from the population of DJs, (2) inputting texts in ODs in VQ, (3) obtaining the top-30 likely VLs from each OD, and (4) calculating the commonality values of terms and VLs.

6.2. Result and Discussion. Tables 4 and 5 are the results of the preliminary experiments. Table 4 shows the commonality of randomly selected DJs. The commonality of terms is about twice as high as that of VLs on average. The median, the maximum, and the minimum of commonalities are also higher in those of terms. Moreover, there is a significant difference within comparison with a paired t-test, assuming the equal variances ($t(158) = 11.48, p < 0.01$). In other words, the commonality of VLs is significantly lower than that of terms. However, when we compared the commonalities of selected three themes, the result shows that the commonalities of VLs are 1.7 to 2.4 times higher than those of terms. This result is totally different compared with randomly selected DJs.

This result is caused by the difference between the term and VL distributions. Although the total number of terms is 27,194, the number of unique terms is 4,971, which is less than that of VLs (Table 3). That is, many terms appear more than once. Figures 3 and 4 show the distributions of the numbers of terms and VLs in the corpus. The left graphs of Figures 3 and 4 are the double logarithmic graphs. The numbers of terms and VLs in each DJ are shown on the horizontal axis and the proportion on the vertical axis. However, the probability of the frequency (k) $p(k)$ is small in the portion where k is large, and there are very few k for which $p(k) > 0$. Owing to this, the double logarithmic graph of the distribution of the frequency is weak toward the noise. Accordingly, we added an order plot on the right of Figures 3 and 4, which is equivalent to the cumulative distribution. Terms and VLs are in accordance with $p(k) \propto k^{-\gamma}$, which becomes a power distribution for both terms and VLs. The power index γ of the term distribution is 1.96 (coefficient of determination:

Table 5: Commonalities of terms and VLs of three themes.

	"transportation"		"health"		"personal activity"	
	Terms	VLs	Terms	VLs	Terms	VLs
Commonality	0.152	0.360	0.153	0.264	0.132	0.229

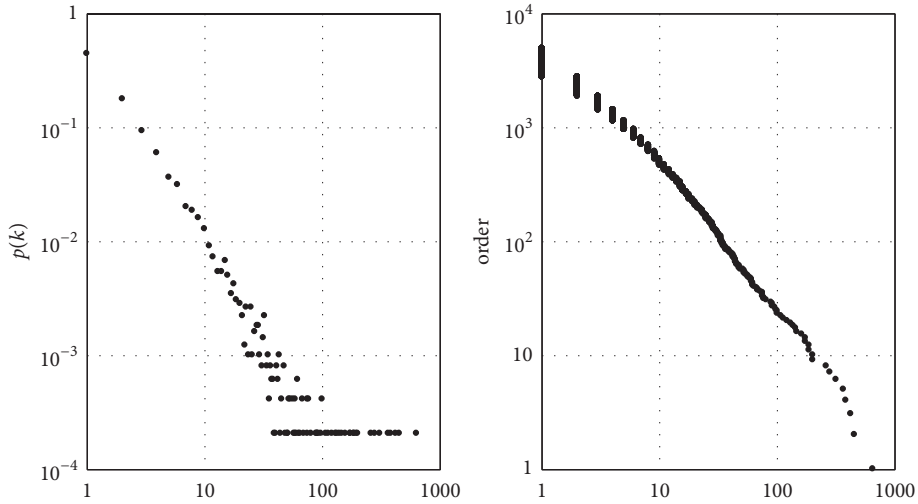


Figure 3: Distributions of the terms in Corpus.

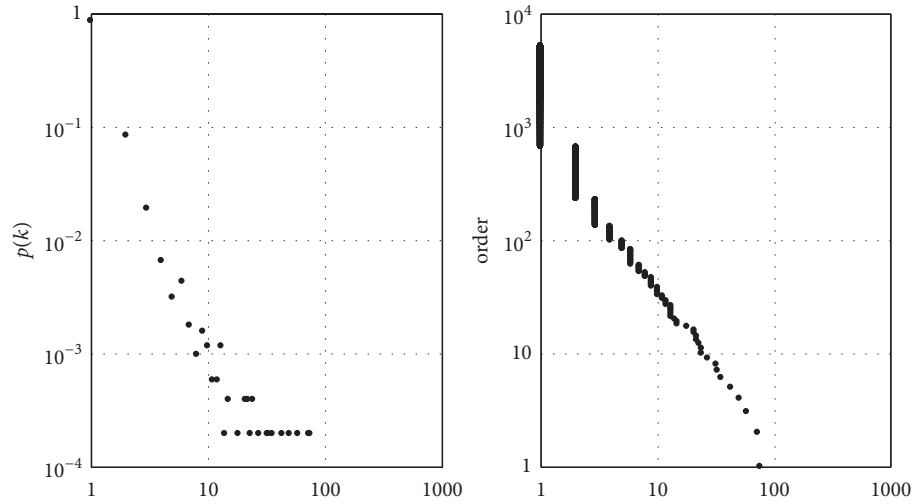


Figure 4: Distributions of the VLs in Corpus.

0.96), and that of the VL distribution is 3.09 (coefficient of determination: 0.72).

The power index of VL distribution is larger than that of terms, which shows that the frequency of many kinds of VLs is rather small. When we randomly choose DJs, the probability of obtaining common VLs using VQ is extremely low. That is, the perspective of the event is so different that we cannot acquire the common VLs. However, by setting the themes, even if the commonality of terms to explain the events is low, the VLs that can be acquired are common to some extent. In other words, the constraints to the perspective of an event by the themes have a function to share data necessary for understanding the event even if the terms expressing the event are somewhat different. For example, if you input the terms "car" and "vehicle" related to the theme "traffic," you can get the relevant VLs "traffic volume" or "location." VQ may be able to bridge the vocabulary gap and obtain the common VLs from different terms when the theme is common.

The purpose of this paper is to recognize the situation of BD in common and to clarify the data (set of VLs) to be acquired. Based on the result and the discussion on the preliminary experiment, it is appropriate to acquire a sentence to explain the BD in the main experiment, obtain VLs from VQ, and compare the commonalities. The same applies to the data on "transportation," "health," and "personal activity"; it

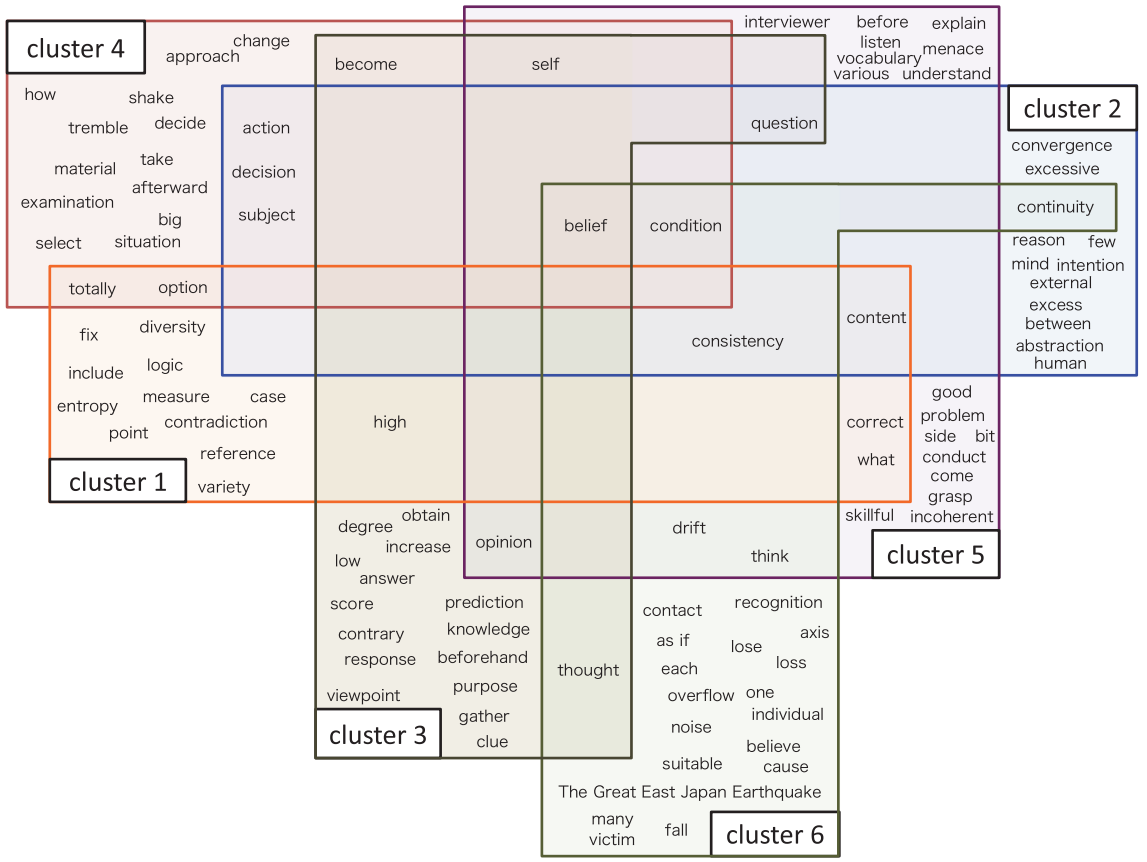


Figure 5: The clustering of terms in sentences about Belief Drift.

can be said that our proposed method works effectively in the situation of BD if the commonality of VLs is higher than the commonality of terms. We can expect the reproducibility of the method in other themes as well. We examine how this result contributes to specifying the data (set of VLs) necessary to explain the situation of DB in the main experiment of the next section.

7. Results

In this paper, we discuss BD by separating engineering researchers and medical researchers (a medical doctor and a psychotherapist). Figure 5 shows the results of morpheme analysis of explanations of BD collected from six researchers. One hundred and five unique terms appear in the sentences in total. Figure 6 shows the frequency of terms shared by more than two researchers. Clusters 1, 2, and 3 are engineering researchers, and clusters 4, 5, and 6 are medical researchers. Table 6 is the number of unique terms in each cluster. Since the numbers are not overly biased, we discuss differences by comparing the number of terms.

We calculate the degree of commonality of elements among researchers by using (10). As a result, the commonality of terms is 0.190 (Table 7). We also compare the commonality of terms between engineering researchers and medical researchers. We define T_{eng} ($T_{eng} = T_1 \cup T_2 \cup T_3$) as the element set of engineering researchers and

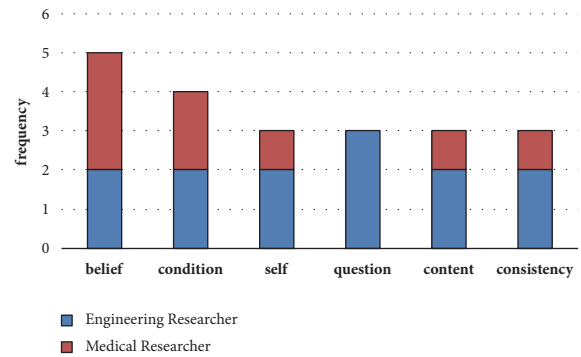


Figure 6: The frequency of terms shared by more than two researchers.

Table 6: Corpus statistics of VQ.

Cluster #	the number of unique terms
1	18
2	20
3	22
4	22
5	28
6	25

Table 7: Commonalities of terms and VLs.

	Terms	VLs
Number of unique elements	105	104
Commonality of all clusters	0.190	0.596
Commonality between clusters of Engineering Researcher and Medical Researcher	0.143	0.519

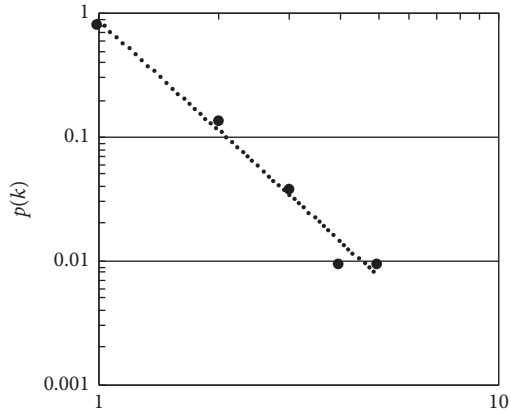


Figure 7: The distribution of terms in the logarithmic graph (the vertical axis represents occurrence probability, and the horizontal axis represents the frequency of terms).

T_{med} ($T_{med} = T_4 \cup T_5 \cup T_6$) as the element set of medical researchers. The commonality is equivalent to a Jaccard index, and it can be represented by commonality(ele) = $|T_{eng} \cap T_{med}|/|T_{eng} \cup T_{med}|$. The commonality of terms between engineering and medical researchers is 0.143 (Table 7). Figure 7 is the distribution of terms in the logarithmic graph.

Conversely, we input texts of each researcher concerning BD in VQ and obtain the top-30 likely VLs. We attained 104 unique VLs in total, which is not very different from the number of terms. Moreover, the similarity for queries of all 30 VLs is higher than 0.110. The maximum is 0.365, and the minimum is 0.113. Two VLs of 0.365 are "the specific scale of renal disease" and "inclusive scale," and nine VLs of 0.113 include "sanitary conditions" and "allergic symptoms."

Figure 8 is the result of the classification of VLs. Some VLs are identified by several phrases such as "non-psychotic psychiatric disorder (symptoms of schizophrenia)" or "ADL score." Note that, since part of the descriptions of VLs protrudes from the frames, VLs are attached with green nodes in Figure 8. Additionally, Figure 9 shows the frequency of VLs shared by more than two researchers. As in Figure 6, clusters 1, 2, and 3 are engineering researchers, and clusters 4, 5, and 6 are medical researchers. Applying (10) to the VL sets, the commonality of VLs is 0.596 (Table 7). Figure 10 is the distribution of VLs in the logarithmic graph.

8. Discussion

Looking at Figure 5, there are relatively few common terms. Only six out of 105 terms are common among over three researchers (Figure 6) and the commonality of terms is 0.190 (Table 7). The terms "belief" and "condition" are relatively common among researchers, but many terms appear only once. The power exponent of the frequency of terms is $\gamma = 2.94$ (the determination coefficient is 0.980) in Figure 7. The frequency of terms follows $p(k) \propto k^{-\gamma}$, which shows that the distribution of terms is a power distribution. The results show that common terms are rarely used to explain BD. That is, there is no commonality among researchers to explain BD.

On the other hand, compared with the commonality of terms in sentences explaining BD, the commonality of VLs is 0.596, which is 3.14 times higher (Table 7). Additionally, fifteen out of 104 VLs are common among more than two researchers (Figure 9), which is larger than for terms. This result is the same as the result when giving the theme as the constraint in the preliminary experiment. Figure 10 shows that the distribution of VLs is not a power distribution, but is the exponential distribution ($p(k') = 2.25 \exp(-1.14k')$ and the determination coefficient is 0.883). This result means that, in the distribution of VLs, the extremely low frequent elements do not occupy the majority as much as the power distribution. Although there are low frequent elements in the exponential distribution, the degree to which the low frequent elements are dominant over the whole is smaller when compared to the power distribution. It shows that VLs important for understanding BD are common among researchers compared with the terms. We also compare the commonality of VLs between engineering researchers and medical researchers. Accordingly, the commonality of VLs between engineering and medical researchers is 0.519, which is 3.62 higher than that in the terms (Table 7). The commonality of the terms in the sentences to explain BD is low not only among all the clusters, but also between engineering researchers and medical researchers. However, when considering VLs, the number of common VLs increases and the commonality is higher than terms.

Considering both the commonalities and distributions, the results suggest that, even if the terms used to explain the state of BD differ, the data (sets of VLs) to be acquired to understand BD are common to some extent. The higher commonality of VLs is similar to the result of giving the theme as the constraint in the preliminary experiment. As we expected, the reproducibility of the method can be guaranteed when the themes are given as the constraints when we use VQ to extract the common data to be acquired to understand the certain situations. Thus, the result supports Hypothesis 1: "even if the agents have different background knowledge in understanding BD, important data for understanding BD is common." Moreover, although it is possible that the vocabulary gap may interrupt a discussion on data necessary for decision making, VQ may be able to bridge the vocabulary

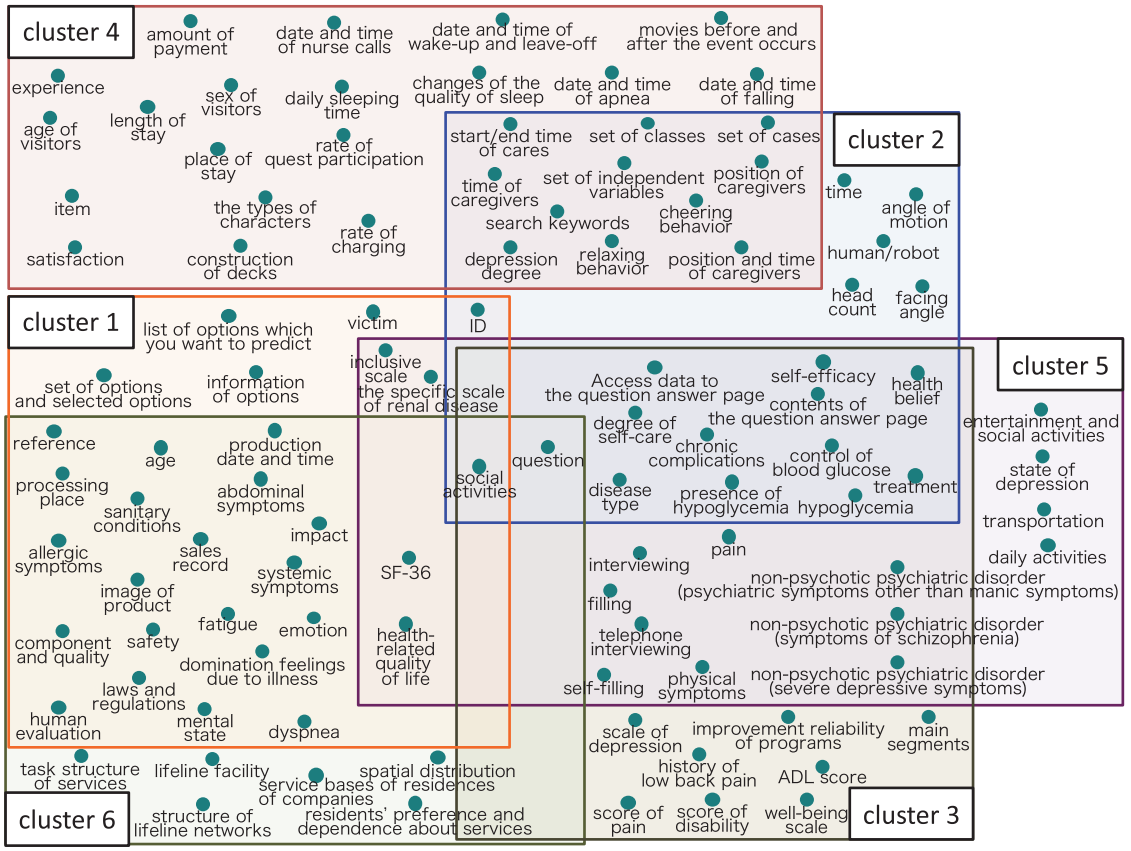


Figure 8: The clustering of VLs inferred from the terms to explain Belief Drift.

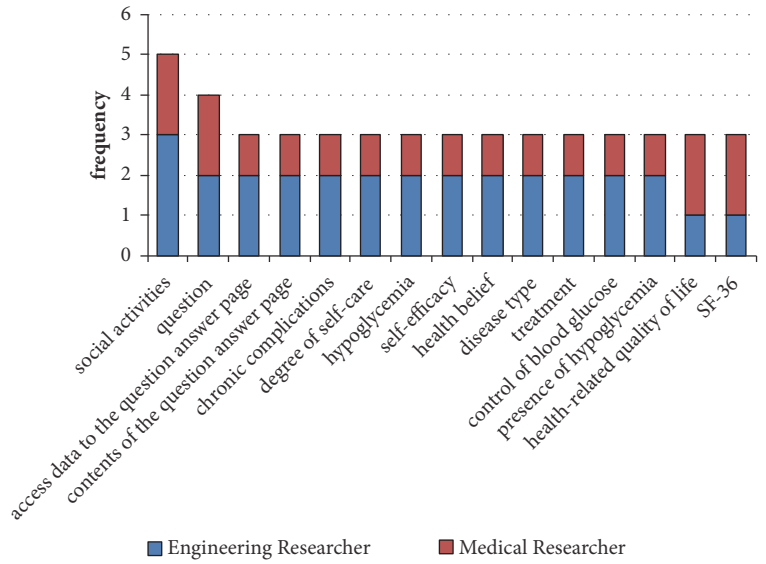


Figure 9: The frequency of VLs shared by more than two researchers.

gap and infer related VLs even if the terms used to explain the common event differ because of the various background knowledge.

The interesting point of the results is that VQ suggests "social activities," "question," "access data to the question answer page," and "contents of the question answer page" which are most common among researchers. In the research project of BD, we are analyzing text data of the question answer site to evaluate the degree of BD. The results of this paper strongly support the hypothesis that the text analysis

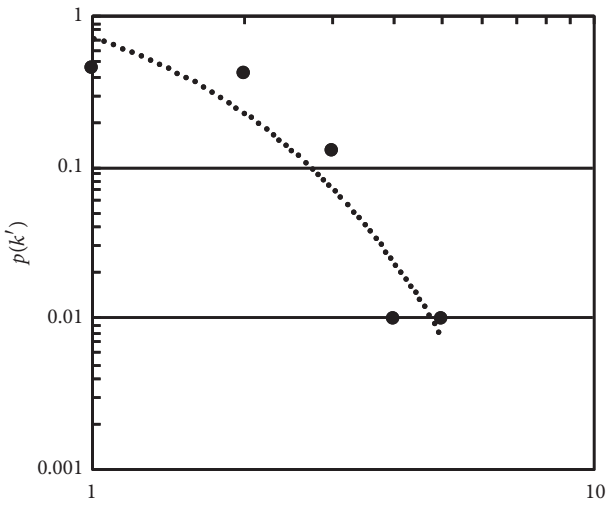


Figure 10: The distribution of VLs in the logarithmic graph (the vertical axis represents occurrence probability, and the horizontal axis represents the frequency of VLs).

of the question answer page may be useful for evaluating and detecting the state of BD.

9. Conclusion

9.1. Summary. Agents with different background knowledge may use different words to describe common concepts. Alternately stated, despite the same framework of thought or understanding, vocabulary gaps may occur when explaining common events or problems such as Belief Drift. Vocabulary gaps may make it difficult to form a common recognition for action, i.e., the acquisition of data or analysis. To bridge these gaps, we tried to extract sets of Variable Labels from different terms explaining the common concept using VARIABLE QUEST. Consequently, although the commonality of terms was low, the commonality of Variable Labels was higher. In other words, even though the terms used to explain events or problems differ, since the framework of thought and understanding are relatively the same, the Variable Labels necessary for understanding the state of BD attained higher commonality.

However, in order to obtain the same performance as these results, a certain amount of information on data (DJ in this paper) is required. It can be said that it is difficult to get satisfactory results if the corpus data is minimal. Also, it is necessary to test the selection protocols of themes given as constraints. In this paper, we did not discuss how much data is reasonable for obtaining results. For our future study, it is necessary to clarify the appropriate number of terms and VLs to obtain sufficient results.

9.2. Future Work. The essential mission of our project is to establish the fundamental methodology and systems providing appropriate information for people whose beliefs are drifting. The result of this experiment suggests information about variables and data we need to acquire to understand situations of BD caused by various factors. In the next stage of our project, we will obtain data according to the advice of medical researchers. Moreover, we will integrate our result with the suggestion obtained from the text analysis of the question answer page.

Data Availability

Data from Data Jacket with Variable Labels and the results used to support the findings of this study have been deposited in the Data Jacket Store repository stored in RDF/XML (http://160.16.227.37/sparql) and are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was partially supported by JST-CREST Grant Number JPMJCR1304 and JSPS KAKENHI Grants Numbers JP16H01836 and JP16K12428. We thank research members of JSPS KAKENHI Grant Number JP16H01836 for proposing the definition of Belief Drift, which worked as the initial input data for the study of this paper. We also appreciate that the tool VARIABLE QUEST used in this paper is based on the presented paper and the fruitful discussion in the MoDAT workshop in IEEE International Conference on Data Mining Workshops (ICDMW) 2017.

References

[1] "Official website of the Department of Homeland Security," https://www.dhs.gov/topic/disasters.

[2] "Official website of the Department of Energy," https://www.energy.gov/.

[3] "Basic Policy and Action Plan for Building IT Disaster-Management Lifeline," in *Proceedings of the IT Disaster-Management Lifeline Promotion Conference in Japan*, 2012, https://japan.kantei.go.jp/policy/it/_full.pdf.

[4] Ministry of Internal Affairs and Communications, "Japan's International Contribution in the Field of ICT for Disaster Management," http://www.soumu.go.jp/menu_seisaku/ictseisaku/bousai_ict/eng/.

[5] Infrastructure Ministry of Land, "Disaster Prevention Portal," http://www.mlit.go.jp/river/bousai/olympic/en/index.html.

[6] M. Tsubokura, S. Kato, S. Nomura et al., "Absence of internal radiation contamination by radioactive cesium among children affected by the fukushima daiichi nuclear power plant disaster," *Health Physics Journal*, vol. 108, no. 1, pp. 39–43, 2015.

[7] M. Tsubokura, S. Kato, S. Nomura et al., "Reduction of high levels of internal radio-contamination by dietary intervention in residents of areas affected by the Fukushima Daiichi nuclear plant disaster: A case series," *PLoS ONE*, vol. 9, no. 6, 2014.

[8] M. Tsubokura, K. Hara, T. Matsumura et al., "The immediate physical and mental health crisis in residents proximal to the evacuation zone after Japan's nuclear disaster: An observational pilot study," *Disaster Medicine and Public Health Preparedness*, vol. 8, no. 1, pp. 30–36, 2014.

[9] Y. Nara, "A Cross-cultural Study on Trust and Risk Perception among Japan, China, and the United States: Focusing on Earthquakes and Nuclear Power Plant Accidents," in *Proceedings of the 16th International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, pp. 1609–1621, 2012.

[10] K. B. Moysich, R. J. Menezes, and A. M. Michalek, "Chernobyl-related ionising radiation exposure and cancer risk: an epidemiological review," *The Lancet Oncology*, vol. 3, no. 5, pp. 269–279, 2002.

[11] H. Nakada, N. Murashige, T. Matsumura, Y. Kodama, and M. Kami, "Informal network of communication tools played an important role in sharing safety information on H1N1 influenza vaccine," *Clinical Infectious Diseases*, vol. 51, no. 7, pp. 873-874, 2010.

[12] A. Sugimoto, S. Krull, S. Nomura, T. Morita, and M. Tsubokura, "The voice of the most vulnerable: Lessons from the nuclear crisis in Fukushima, Japan," *Bulletin of the World Health Organization*, vol. 90, no. 8, pp. 629-630, 2012.

[13] Y. Ohsawa, N. Kushiro, M. Hirano, M. Tsubokura, and M. Kami, "Technologies for Generating and Providing Information to Suppress Belief Drifts and Reinforce Psychological Resilience," in *a Project of JSPS KAKENHI Grant Number JP16H0*, https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-16H01836/.

[14] M. Boisot and A. Canals, "Data, information and knowledge: Have we got it right?" *Journal of Evolutionary Economics*, vol. 14, no. 1, pp. 43–67, 2004.

[15] T. Hayashi and Y. Ohsawa, "VARIABLE QUEST: Network Visualization of Variable Labels Unifying Co-occurrence Graphs," in *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 577–583, New Orleans, LA, November 2017.

[16] Y. Hayashi, K. Miwa, and J. Morita, "A laboratory study on distributed problem solving by taking different perspectives," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 333–338, 2006.

[17] J. S. Metcalfe, *Evolutionary Economics and Creative Destruction*, Routledge, UK, 1998.

[18] Y. Ohsawa, H. Kido, T. Hayashi, and C. Liu, "Data jackets for synthesizing values in the market of data," *Procedia Computer Science*, vol. 22, pp. 709–716, 2013.

[19] Y. Ohsawa, H. Kido, T. Hayashi, C. Liu, and K. Komoda, "Innovators marketplace on data jackets, for valuating, sharing, and synthesizing data," in *Knowledge-Based Information Systems in Practice*, vol. 30 of *Smart Innovation, Systems and Technologies*, pp. 83–97, Springer International Publishing, Cham, 2015.

[20] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web*, vol. 4825 of *Lecture Notes in Computer Science*, pp. 722–735, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[21] T. Hayashi and Y. Ohsawa, "Matrix-based method for inferring variable labels using outlines of data in data jackets," in *Advances in Knowledge Discovery and Data Mining*, vol. 10235 of *Lecture Notes in Computer Science*, pp. 696–707, Springer International Publishing, Cham, 2017.

[22] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[23] T. Hayashi and Y. Ohsawa, "Knowledge structuring and reuse system design using RDF for creating a market of data," in *Proceedings of the 2nd International Conference on Signal Processing and Integrated Networks, SPIN 2015*, pp. 607–612, India, February 2015.

[24] T. Kudo and Y. Matsumoto, "Japanese dependency structure analysis based on support vector machines," in *Proceedings of the the 2000 Joint SIGDAT conference*, pp. 18–25, Hong Kong, October 2000.

[25] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

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 Neutral	Factor 2 Frank	Factor 3 Dialect	Factor 4 Polite	Factor 5 Feminine	Factor 6 Crude	Factor 7 Aged	Factor 8 Interrogative	Factor 9 Approval	Factor 10 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.

Hindawi
Advances in Human-Computer Interaction
Volume 2018, Article ID 6780564, 12 pages
https://doi.org/10.1155/2018/6780564

Research Article

Emergentist View on Generative Narrative Cognition: Considering Principles of the Self-Organization of Mental Stories

Taisuke Akimoto

Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Correspondence should be addressed to Taisuke Akimoto; akimoto@ai.kyutech.ac.jp

Received 25 June 2018; Accepted 31 October 2018; Published 12 November 2018

Guest Editor: Akinori Abe

We consider the essence of human intelligence to be the ability to mentally (internally) construct a world in the form of stories through interactions with external environments. Understanding the principles of this mechanism is vital for realizing a human-like and autonomous artificial intelligence, but there are extremely complex problems involved. From this perspective, we propose a conceptual-level theory for the computational modeling of generative narrative cognition. Our basic idea can be described as follows: stories are representational elements forming an agent's mental world and are also living objects that have the power of self-organization. In this study, we develop this idea by discussing the complexities of the internal structure of a story and the organizational structure of a mental world. In particular, we classify the principles of the self-organization of a mental world into five types of generative actions, i.e., connective, hierarchical, contextual, gathering, and adaptive. An integrative cognition is explained with these generative actions in the form of a distributed multiagent system of stories.

1. Introduction

The computational modeling of an autonomous intelligence that can adapt to external environments (physical and social situations including other humans) is an essential issue for realizing human-like artificial intelligences. In cognitive architecture studies, computational frameworks of autonomous intelligence have been explored with biological inspirations, including psychology and neuroscience [1, 2]. In the early years of artificial intelligence, Schank and his colleagues argued the importance of narrative ability and narrativity-based memory in higher-level cognition and learning. They proposed several significant theories, including script knowledge [3] and a dynamic memory framework [4]. His dynamic memory framework demonstrated a systematic cognitive mechanism of flexible reminding (remembering), reconstruction, generalization, and organization of story-form memories. Although a large part of his idea was not implemented, it provided an important insight into the autonomous development of intelligence.

Based on the above background, we assume that generative narrative cognition is an essential aspect of an autonomous intelligence, which develops through interactions with external environments. Here, generative narrative cognition refers to an agent's mental system of dynamically generating and organizing stories for interacting and adapting to environments. In this study, we use the term "story" to refer to a mental representation of a part of the world of an agent. It is used as a concept unifying episodic memories, autobiographical memories, the contextual structures of current situations, prospective memories, planned or imagined futures, and fictive or virtual stories. On the contrary, a narrative expressed through language or other media of expression is referred to as a "narrative" or a "discourse."

We can state several reasons for the importance of generative narrative cognition in cognitive architectures. First, a story is a universal information format that integrates various informational elements, including events, entities, relationships, abstract concepts, intents, goals, emotions, nonverbal information (e.g., memories of visual images), and

hypothetical events. Second, a narrative is a universal way of communicating world information with others. Third, a story forms the contextual structure of an unfolding situation involving temporal reach into the past (i.e., experiences and results of one's actions) and future (i.e., expectations and plans). In this sense, a story is the basis of a higher-level perception-action system. Fourth, memories of past experiences become reusable knowledge when they are organized as stories. Moreover, the importance of narrative ability in cognitive architectures, artificial agents, and human-computer interaction has been discussed from various perspectives [5–8].

However, the computational modeling of generative narrative cognition is an extremely complex problem that has challenged researchers for many years in artificial intelligence studies [9, 10]. Although most previous narrative generation systems have focused on the production of narrative texts such as fairy tales and literary narratives, the basic problem is common: using generative narrative cognition as the foundation of an agent's mind. There exist several difficult problems in the computational modeling of generative narrative cognition. In particular, a story or a narrative has a complex structure, and human narrative cognition is based on a vast store of experiential knowledge, including informal, tacit, and cultural knowledge. Such a complex problem is difficult to model on the basis of classical symbolic processing. Although connectionist models including deep neural networks are applied to various domains including image-recognition systems and end-to-end natural language processing systems, this type of approach does not fit the essential part of generative narrative cognition. The critical issue of generative narrative cognition is to explore a computational framework and the underlying principles of the generation and organization of stories in the mental system of an agent. Therefore, we must seek an alternative method from a long-term perspective.

In this paper, we propose a conceptual-level theory for the computational modeling of generative narrative cognition based on a type of emergentist approach. Our basic idea can be described as follows: stories are living objects having the power of self-organization. This is similar to a multiagent system. Here, an agent is not a character, but a representational element in an agent's mental system. For example, in Minsky's *The Society of Mind* [11], the cognitive mechanism of a mind is explained as a type of distributed multiagent system based on the collaborative activities of diverse simple functional agents. However, we assume that the central agents forming a mind are stories.

The rest of this paper is organized as follows. Sections 2 and 3 describe the basic idea of how generative narrative cognition plays a crucial role in an autonomous intelligence. Section 4 discusses the necessity of an emergentist approach for the computational modeling of generative narrative cognition. Based on this idea, Section 5 provides a macroscopic classification of the principles of the self-organization of a story and a mental world formed by many stories. Section 6 contains concluding remarks with future research directions. Although this paper provides only conceptual descriptions, creating a vision for solving this complex problem (generative narrative cognition) is a significant step for the future of artificial intelligence.

2. Stories Forming an Agent's Mental World

The motivation behind this study is based on an assumption that generative narrative cognition is an essential aspect of an autonomous artificial intelligence. From this perspective, we have been addressing the conceptual systematization of a cognitive architecture. The initial concept is presented in [12]. The key concept of our architecture is an agent's mental world formed as an organization of many stories.

2.1. Story. In general terms, a story refers to the information of chronologically and semantically organized events recounted in a narrative. Here, an event refers to a character's action or a happening (e.g., "Taro eats an apple").

The notion of story is rooted in narratological terminologies (narratology is the discipline of theoretical studies on narrative, inspired by structuralism and semiology). In terms of narratology, a "narrative" basically refers to an expression of events in a real or fictional world based on a language or other sign system [13]. However, a narrative has a close relationship with the form of mental representation of knowledge and memory. To clearly distinguish the representational aspect from an expressed narrative, we introduce the notions of story and discourse based on a reinterpretation of narratological terminology [13, 14]. The terms "story" and "discourse" are generally used to distinguish between the content and expression planes of a narrative. More precisely, a discourse refers to the narrative text itself and a story corresponds to the content, i.e., information of events recounted in a discourse or a text. However, because a story is intangible, the notion of stories is slightly unclear. From a narrative-communication perspective, the relationship between the content and expression planes of a narrative can be reinterpreted as the relationship between a mental representation and the surface expression. A sender (author or teller) writes or tells a discourse based on a story that is remembered or generated inside the mind. A receiver (reader or hearer) mentally constructs a story by interpreting or comprehending the discourse. Stories between the sender and the receiver are not the same objects.

Based on the above conception, we use the term "story" as a uniform mental representation involving an episodic memory, an autobiographical memory, the contextual structure of a current situation, a prospective memory, a planned or imagined future, and a fictional or virtual story.

2.2. Mental World. An agent's mental world contains individual meaning and a rich temporal extent with numerous and diverse stories, as illustrated in Figure 1. A story corresponds to a piece of the world for an agent. Stories contained in a mental world can be classified from several perspectives. In the relationship with an external world, these stories include past, future, and fictive or hypothetical stories. With respect to the manner of generation, there are stories based on an agent's own experiences (experience-based stories),

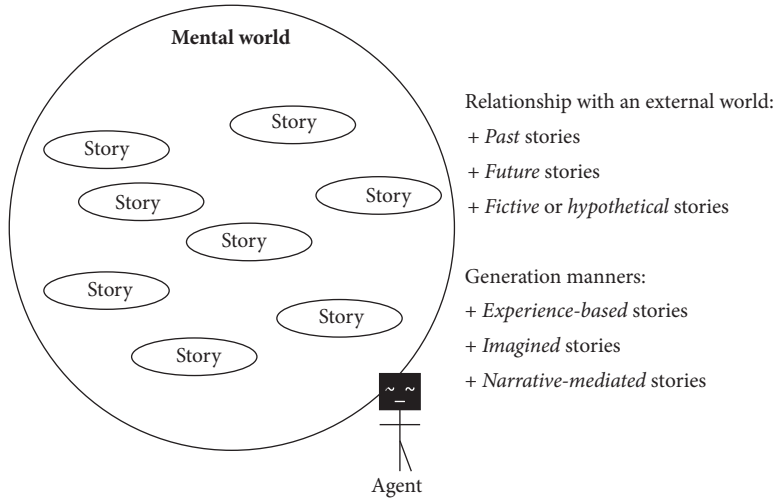


Figure 1: An agent's mental world formed by stories.

imaginative power (imagined stories), and interpretation or comprehension of others' narratives (narrative-mediated stories). A more systematic description of structural properties and functions of an agent's mental story are presented in our previous study [15].

From a perspective of computational knowledge representation, a story itself forms only a relational structure of concrete events and entities. However, we assume that the semantic aspect of a story is underpinned by associations with the following three types of mental elements:

(i) **Concept:** a concept corresponds to a primitive linguistic element corresponding to meanings of a word. Concepts include general concepts (i.e., nominal, verbal, adjectival, and adverbial concepts) and ontological (or proper) concepts for identical entities.

(ii) **Schema:** a schema refers to a generalized structure based on one or more concrete stories. A schema is a structured composition of two or more concepts or (sub)schemas. Schemas underpin top-down and abstract-level cognitions of stories. The idea of schemas is rooted in Minsky's frame theory [16]. Script knowledge [3] and memory organization packets (MOPs) [4] form the schematic knowledge relevant to narrative cognition.

(iii) **Mental image:** a mental image represents nonsymbolic information, including verbal, visual, auditory, and haptic images.

3. Dynamic Generation and Organization of Mental Stories

How generative narrative cognition plays an essential role in an agent's intelligence can be explained from its functional generality. In particular, generative narrative cognition forms the common basis of a higher-level perception-action system, linguistic communication about world information, and formation of a self or identity. We will describe the first two aspects later in this section. The thoughts behind the third aspect (formation of a self or identity) can be found in philosophy, psychology, or other disciplines [17–19].

In an autonomous intelligence, dynamic generation and organization of stories includes two aspects, i.e., (1) interaction with an environment by generating a story and (2) adaptation to environments by developing the organizational structure of a mental world. Figure 2 illustrates these two aspects, and we will describe each aspect in the following two subsections. Here, assumed environments include various social and physical situations over the course of an agent's activities, e.g., shopping at a supermarket, climbing a mountain, linguistic communication with others, housework, and creation of literary work.

3.1. Interaction with an Environment by Generating a Story. An agent interacts with an environment based on a story. This idea is rooted in our previous consideration of the structure of an agent's subjective world while interacting with an environment [15]. A story for interacting with an environment can be explained through the following two perspectives:

(i) **Action and perception:** in an agent's mind, interaction with an environment is based on the continual (re)construction of a story (see Figure 3(a)). In particular, acting in an environment corresponds to performing mentally constructed events placed in the future. Perceiving the movement of an environment, including the results of one's actions, corresponds to the construction of past episodic events. Both actions and perceptions always occur in the context of a story, i.e., a chain of events across the past (experiences and results) and future (expectations and plans).

(ii) **Expression and interpretation (linguistic communication):** a narrative is the universal way of exchanging world information, and story generation is the core mental process in this activity (see Figure 3(b)).

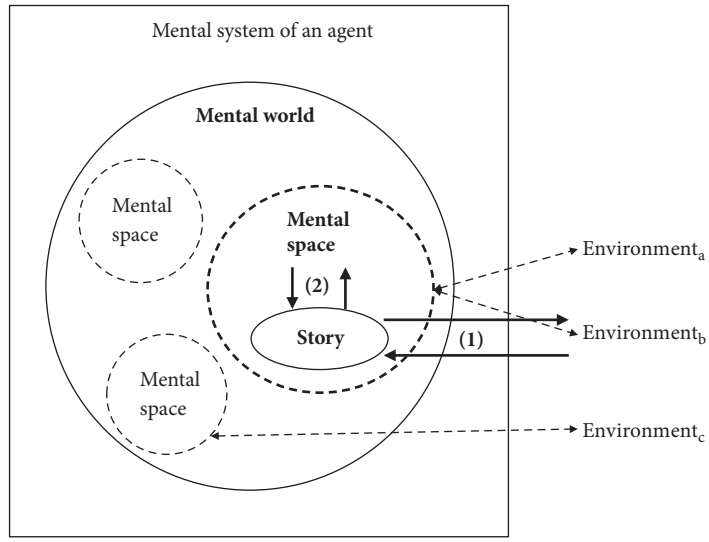
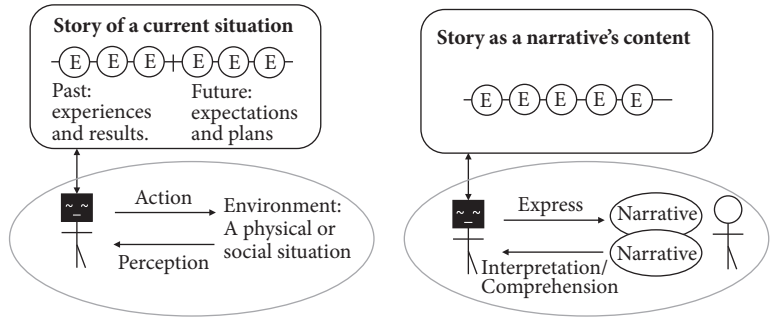


Figure 2: Two dynamic aspects of a mental world: (1) the generation of a story and (2) development of the organizational structure of a mental world.



(a) Action and perception (b) Linguistic communication

Figure 3: Action/perception and linguistic communication based on a story.

When an agent expresses a narrative to others, a story for the source of the narrative is remembered or imagined in the mind. The mental process of recalling includes not only memory retrieval but also flexible editing of memories (stories) according to the situation or goal of the narrative communication. On the contrary, the fundamental mental process of interpreting or comprehending another's narrative is to compose a mental story over the course of hearing/reading it.

It is also important that generative narrative cognition integrates the above two fundamental human activities. A human develops mental stories in various ways, e.g., the organization of one's own experiences, imagination, and the reception of various narratives. The acquisition of stories from narratives has great significance in constructing rich world knowledge beyond one's own experiences and imaginative power. In addition, humans cocreate world knowledge by communicating narratives, such as the histories and current states of societies and visions for the future.

3.2. Adaptation to Environments by Developing a Mental World. A mental world is a holistic memory system that provides knowledge resources for generating new stories. At the same time, a mental world organizes the pieces required to direct the dynamic generation of stories in various environments. The cognitive mechanism of the autonomous development of a mental world through the accumulation of experiences is the foundation of an environmental adaptability, i.e., the potential to build abilities to generate adequate stories in various environments.

The developmental process of a mental world is conceptualized as the formation of a "mental space," which is the basic organizational unit of a mental world. This concept is inspired by Schank's dynamic memory framework [4]. In general terms, a mental space corresponds to a generalized structure of similar stories. The basic role of a mental space is to provide a framework that directs and restricts story generation for smoothly interacting with similar type of environments, e.g., shopping in a supermarket and communication about a specific theme (politics, local events, etc.). In addition, a mental space organizes memories of experiences,

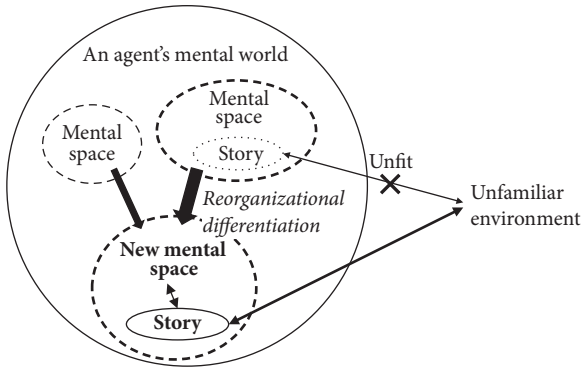


Figure 4: Reorganizational differentiation of a mental space.

i.e., previously generated stories in that space, as knowledge resources for generating a new story in a similar situation.

The framework of a mental space is formed by a schema, i.e., a general structure covering similar type of stories. In particular, the schema of a mental space provides a thematic and stylistic framework based on a compound of multiple subschemas. A schema adaptively forms or reforms by the generalization of stories and reorganization of existing schemas. When an agent is facing a relatively familiar environment, a story is generated in an existing mental space corresponding to that environment. The schema of this space may be adjusted or reformed according to a generated story (i.e., experience). Moreover, when an agent is facing an unfamiliar environment, the agent's mental system tries to adapt to that environment by reorganizing one or more existing schemas to create a story with a new mental space (see Figure 4). We call this specific mental process of adapting to a new environment "reorganizational differentiation." (This notion is similar to abduction in C. S. Perce's theory.)

4. Necessity of an Emergentist Approach to Generative Narrative Cognition

Computationally implementing a dynamic mental world described in the previous section is a hugely complex problem. For solving this problem, we argue the necessity of an emergentist approach.

4.1. Structural Complexity of a Story. The necessity of an emergentist approach can be described from the following two perspectives:

First, from a cognitive perspective, the developmental process of a mental world (or a memory system in general terms) is generally assumed as a type of self-organization phenomenon. The ability to build diverse mental spaces with new intellectual functions is a key aspect of a dynamic mental world. These functions should not be externally embedded but emerge through adaptive interactions with environments.

The second reason refers to the structural complexity of a story itself. A story as a world representation can be viewed as a complex structural object in which events (a character's action or a happening) and entities (an individual existence including a character, object, and place) are organically organized. Although the central elements of a story are events, these events are potentially accompanied by various types of informational elements, e.g., relationships, abstract concepts, intents, goals, emotions, nonverbal information, and hypothetical events. Because of this property, it is assumed that the cognitive process of story generation is based on a flexible collaboration of multiple cognitive modules. In addition, the structure of a story involves interdependencies in the whole-part and part-part relationships. In a well-organized story, for example, a small change in a story's part may cause incoherence in the story's whole structure or the lack of contextual coherence over the course of its events. Based on an extensional reinterpretation of our hierarchical graph model of multidimensional narrative structure [20] and a structural conception of an agent's subjective world [15], we arrange general structural properties of a story as follows (Figure 5 illustrates these notions):

(a) A story has a hierarchical whole-part structure. A higher-level part (e.g., a scene, a semantic or functional unit, and a larger section) of a story is formed from two or more lower-level parts (e.g., events and scenes).

(b) Parts of a story (including events) are mutually connected by temporal, causal, or other types of relationships.

(c) There are two types of interdependencies in a story's hierarchical structure: (1) vertical interdependencies, including top-down restriction (from a higher-level part to the lower-level parts) and bottom-up abstraction (from lower-level parts to the higher-level part) and (2) horizontal interdependencies based on the contextual coherence between parts.

(d) A story contains a story world, i.e., the organization of entities relevant to the story. It is similar to the setting of the world and there exists an interdependency between a story and the story world.

(e) A story is formed based on the reconstructive reuse of existing mental resources including relationships with other stories. In this sense, a story itself is not an independent mental object.

(f) There exists an interdependency between a story and an external environment. In particular, a story based on one's own experiences is formed based on perceived environmental information. At the same time, the expectational aspect of the agent's story directs or restricts the perception of environmental information.

In the above six items, (e) is derived from the discussion in Section 3.2 and (f) is an ordinary notion from a cognitive perspective. To explain notions (a)–(d), the next subsection uses an example story structure.

4.2. Example Story Structure. Box 1 shows an English expression of a simple story written by the author. An example

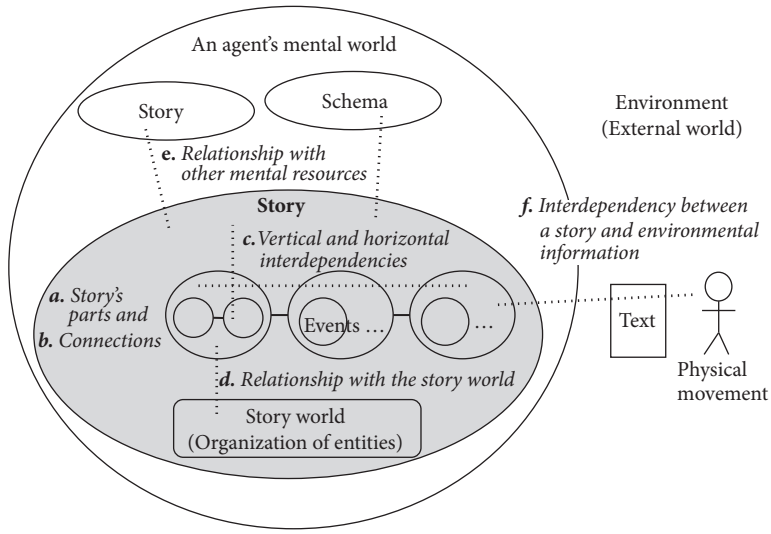


Figure 5: Structural complexity of a story. Dot-lines represent interdependencies or relationships among a story's parts and other mental objects.

(s1) Lisa and Sally are sisters that get along well. (s2) They play together every day. (s3) One day, Sally found a chocolate in Lisa's desk when Lisa was not in the room. (s4) Sally stole and ate the chocolate. (s5) The next day, Lisa saw that the chocolate was gone. (s6) Lisa assumed that the chocolate was eaten by Sally. (s7) Then, Lisa threw Sally's doll out a window. (s8) Sally cried.

Box 1: English expression of an example story.

of how the structure of this story text can be interpreted is shown in Figure 6. In this structural representation, each sentence in Box 1 is simply imagined as an event. From the perspectives of (a)–(d) in the above list, we can examine the cognitive processes of creating and manipulating this fictional story.

(i) **Hierarchical structure (a):** this story is divided into three scenes (intermediate parts): the setting (daily life), the stealing by Sally, and the revenging of the theft by Lisa.

(ii) **Relationship between parts (b):** there are anteroposterior and causal relationships between the different parts. The anteroposterior relationships are depicted as arrowed lines. The block arrow from the second scene to the third scene represents the causal relationship, i.e., Lisa's reason or motivation for taking revenge.

(iii) **Vertical interdependency (c):** a change in a part may cause a change in the higher- and/or lower-level parts. For example, when event s4 is changed to "Sally put a cookie beside the chocolate," it will propagate and influence the meaning of the second scene, e.g., "gift to Lisa."

(iv) **Horizontal interdependency (c):** a change in a part also propagates in horizontal directions. For example, when the second scene is changed to "gift to Lisa," the third scene, "revenging of the theft by Lisa," becomes an unnatural reaction (at least from our commonsense perspective). This inconsistency causes a global reorganization of a story.

(v) **Relationship with the story world (d):** in the structural representation in Figure 6, the story world contains several entities, i.e., Lisa, Sally, a desk, a chocolate, and a doll, and their relationships such as "Lisa and Sally are sisters" and "Lisa and Sally like chocolate." A change in the setting of the world propagates to events or the story's parts. For example, if the setting is changed to a fantasy world (e.g., Lisa is a witch and all children like dried frogs), the chocolate element will become a dried frog and the way Lisa takes revenge will change to become a magical attack.

4.3. Emergentist Approach. In studies of computational narrative generation, an orthodox approach is to model the process of generating a story (or a narrative) based on a centrally controlled symbolic processing (see Figure 7(a)). For example, narrative generation using a story grammar is a traditional story-generation method [21, 22]. The simplest implementation is a pipelined procedure of composing a story using a story grammar in a top-down manner, from the abstract structure to the detailed contents. However, this type of inflexible framework will be limited when

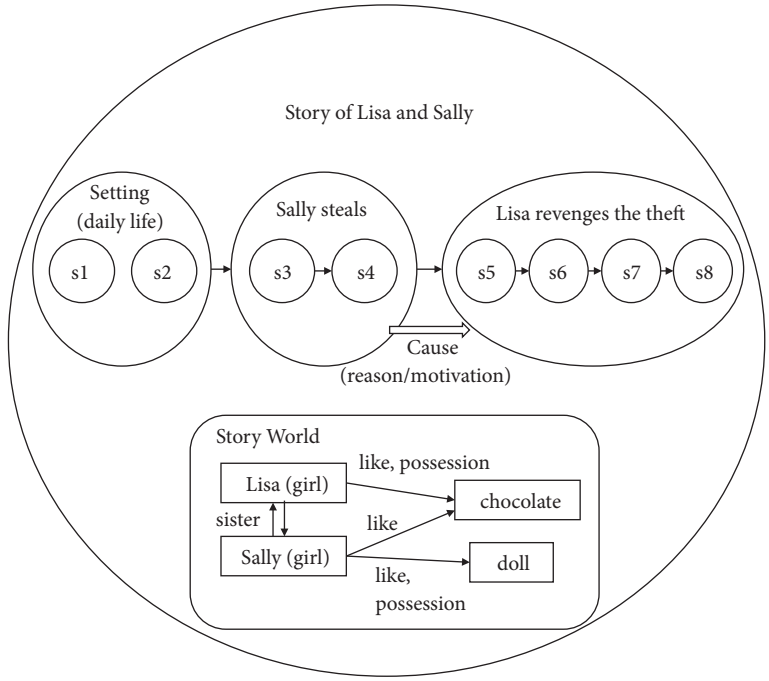
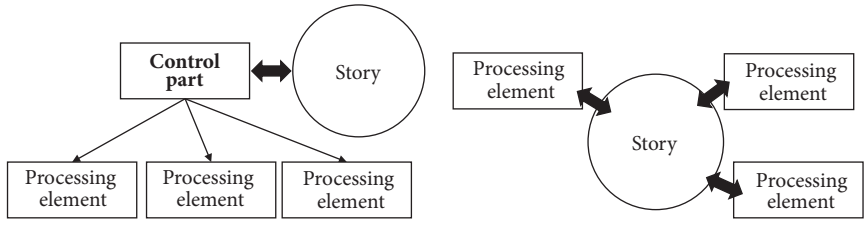


Figure 6: A structural representation of the example story in Box 1 (s1–8 in the small circles correspond to sentences in Box 1).



(a) Centrally controlled architecture (b) Blackboard architecture

Figure 7: A centrally controlled architecture and a blackboard architecture.

modeling a complex and dynamic narrative cognition and does not lead to our objective. An advanced challenge is to model narrative generation as a more flexible procedure. For example, a blackboard architecture (see Figure 7(b)) that shares a processing object, i.e., a story, among various types of cognitive modules has the potential of flexible collaboration of cognitive modules in generating a story (e.g., [23]). However, a blackboard architecture requires knowledge of the principles of directing the collaborative cognitive activities, and how to model it remains a difficult problem.

Therefore, we propose an alternative approach as an extension of the blackboard architecture. Our basic idea is to build the power of self-organization into stories and mental spaces themselves. Figure 8 illustrates this concept. Each part at each level in a story structure has the power of generating one's own structure in the relationship with other mental objects. The story's whole structure emerges from distributed collaborative functioning of the parts (see Figure 8a). On the contrary, a mental space forms one's own schema through interactions with stories and other mental spaces (see Figure 8b). We introduce the term "generative actions" to refer to the basic principles of driving these self-organization activities. A mental world is developed by the distributed collaborative functioning of the generative actions of stories and mental spaces.

5. Classification of the Generative Actions of a Story and Mental Space

Based on the above concept, we can classify generative actions in a mental world. According to the structural properties of a story and the notion of a mental space, we can classify generative actions into five basic types: *connective*, *hierarchical*, *contextual*, *gathering*, and *adaptive*. The last type (adaptive) drives the self-organization of a mental space. The first four types, which drive the self-organization of a story, are derived from the structural complexity of a story; this is described in Sections 4.1 and 4.2. In particular, the connective action is relevant to (b) and (c), the hierarchical action is relevant to (a)

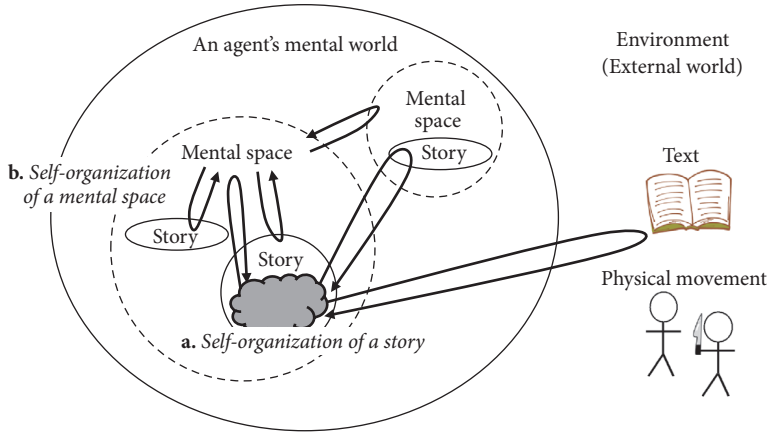


Figure 8: Powers of the self-organization in a mental world.

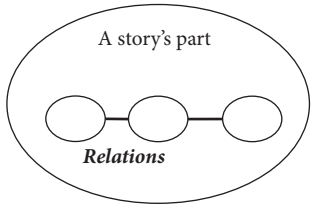


Figure 9: Connective actions.

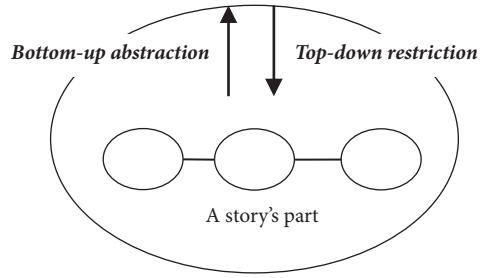


Figure 10: Hierarchical actions.

and (c), and the contextual and gathering actions are relevant to (c)–(f).

In the following subsections, the concept and principles of each basic type are presented. The presented principles are basically rooted in general notions of cognitive science and artificial intelligence studies such as abstraction, generalization, analogy, schema, and relationship. However, the main contribution of the following ideas is to provide a conceptual-level theory of an integrative cognition from an emergentist perspective.

5.1. Connective Actions: Relation. A connective action is the most fundamental action of organizing events. The basic principle of making a connection is a relationship between a story's parts. The major relationship types in a story structure are as follows (see also Figure 9):

(i) **Temporal relationship:** this denotes a relative temporal relationship between two parts. Allen [24] classified temporal relationships between two events or actions into "before," "equal," "meets," "overlaps," "during," "starts," "finishes," and their inverses, on the basis of anteroposterior relationships and temporal intervals of events. This classification can be adopted to temporal relationships in a story.

(ii) **Causal relationship:** this denotes a causal relationship between any two parts of a story.

(iii) **Other type of relationships:** because the computational modeling of relationships in a story structure or narrative is an essential but complex problem, further consideration is required. For example, various types of relationships in a discourse structure are proposed in the area of natural language processing, e.g., Hobbs's coherence relations [25] and the rhetorical structure theory by Mann and Thompson [26].

5.2. Hierarchical Actions: Top-Down Restriction and Bottom-Up Abstraction. Hierarchical actions form or reform upper- and lower-level structures in a story. They are the basis of the vertical interdependency between upper- and lower-level parts. Hierarchical actions can be broadly classified into top-down restriction and bottom-up abstraction (see also Figure 10):

(i) **Top-down restriction:** a higher-level part restricts the lower-level structure by creating the expectation of a blank based on a schema or similar story that is associated with the higher-level part. An expectation refers to the generation of subsequent information in a dynamic story-generation process during interactions with an environment. In particular, an expectation drives the perception of environmental information based on a schema or similar story. A blank denotes a lack of information in the lower-level structure, and it drives the filling of that part using a schema or similar story.

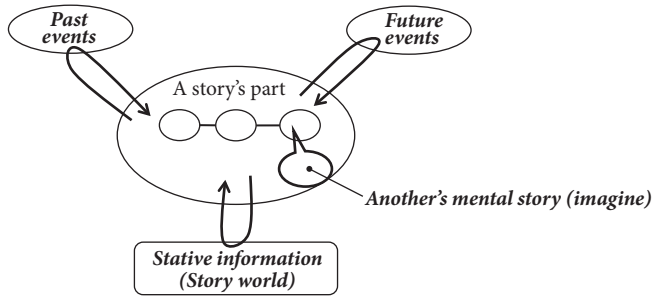


Figure 11: Contextual actions.

(ii) **Bottom-up abstraction:** in this action, a higher-level part is formed or reformed from an aggregation of lower-level parts. This type of cognition, which generates higher-level meaning from lower-level elements, is generally referred to as abstraction [27]. In a story, abstractions arise in various structural levels, i.e., temporal segmentation of events from sensory information and the formation of a higher-level meaning from two or more events or parts. Whereas an abstraction is a bottom-up action, it always functions under top-down restriction. In particular, a higher-level structure (the meaning and power of top-down restriction) is formed by matching and reorganizational diversion of associated mental resources, i.e., a schema or another story's part that is similar to the lower-level structure.

5.3. Contextual Actions: Associations of Events and States. Using contextual actions, a part draws contextual information relevant to the organization of one's own structure. These actions generate horizontal interdependency in a story. Contextual actions can be classified into the following subclasses, from the perspective of the position of source information (see also Figure 11):

(i) **Past:** a part draws relevant past events (e.g., the background or reason behind an action by the agent or another character).

(ii) **Future:** a part draws relevant future events (e.g., an intent, desire, objective, and goal for the agent's action).

(iii) **State:** a part draws relevant stative information from the story world (e.g., a character's emotion and whether conditions).

(iv) **Another's mental story:** a part associates another's imagined mental story. This is relevant to the ability to imagine others' mental states, known as the theory of mind. From a structural perspective, an imagined mental story about another person is represented as a nesting of stories ("a story within a story") [15].

5.4. Gathering Actions: Mental Space, Similarity/Analogy, and Perception. Gathering actions gather mental resources externally (memories including stories, schemas, mental images, and concepts) or environmental information (perception) for generating their own structures of a part. These mental resources are used by other types of generative actions, as materials, general structures, or cases. Gathering actions include the following three subclasses (see also Figure 12):

(i) **Mental space:** a story's part gathers mental resources from the mental space in which the story is generated. As we described in Section 3.2, a mental space organizes relevant knowledge including (sub)schemas and stories for generating stories in similar environments.

(ii) **Similarity/analogy:** similarity or analogy is a key principle for the flexible reuse of existing mental resources across boundaries of mental spaces or problem domains. Particularly, analogy is an essential human cognition for reusing mental resources by making a structural correspondence between two different representational elements [28–30]. Case-based reasoning [31] is also rooted in analogical cognition.

(iii) **Perception:** perception is the action of gathering environmental information. It is basically driven by an expectation as a top-down restriction. When a part gets unexpected information from an environment, reformation of part of the story may be done to maintain coherence.

5.5. Adaptive Actions: Generalization and Differentiation. Adaptive actions form or reform the schema of a mental space to adapt environments. Whereas the above four types are a story's activities, adaptive actions are the activities of a mental space. Adaptive actions can be classified into the following three types (see also Figure 13):

(i) **Inductive generalization:** a mental space forms or reforms one's own schema based on structural commonality among stories in that space. Abstraction of a story (forming a higher-level structure of a story) and analogies between stories (creating structural correspondences between stories) provide the basis for this action.

(ii) **Failure-based generalization:** a mental space adjusts one's own schema according to a failure in interacting with an environment. A failure refers to a type of negative feedback from an environment. A similar concept is discussed in Schank's dynamic memory

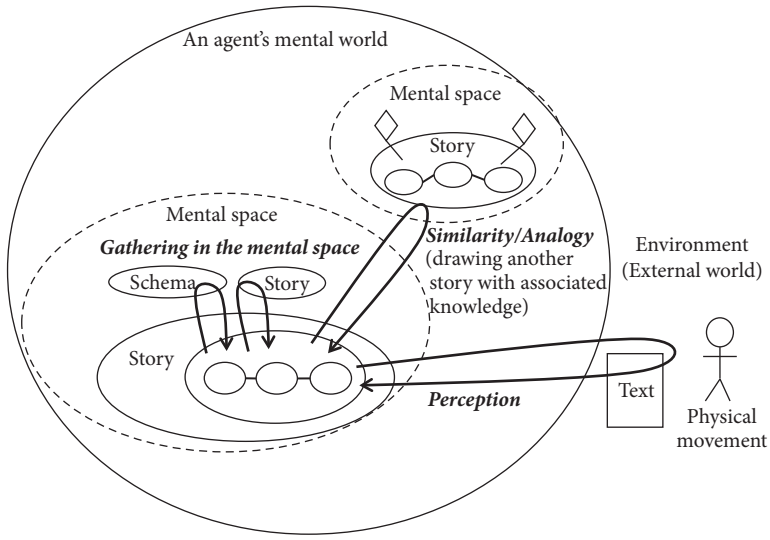


Figure 12: Gathering actions.

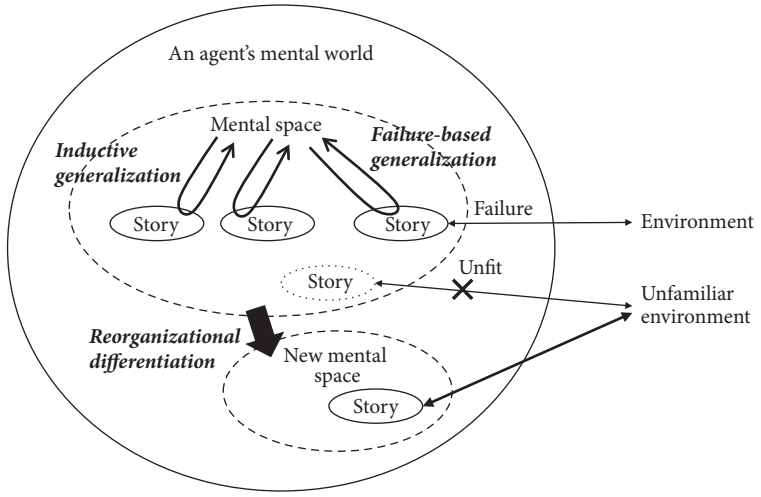


Figure 13: Adaptive actions.

framework [4]. It is also relevant to the general notion of reinforcement learning.

(iii) **Reorganizational differentiation:** when an agent faces an unfamiliar environment that existing mental spaces do not cover, a new mental space emerges in the agent's mental world to generate a story for interacting that environment. This new mental space is formed by the reorganization of existing mental spaces or mental resources. We call this mental action reorganizational differentiation. This is the essence of the ability to develop by adapting to environments. At the same time, this is the fundamental principle of creativity, which seeks to construct new ideas, styles, social environments, etc.

5.6. Integrative View. As we described previously, the key point of the proposed theory is to integrate various types of cognition in the form of a distributed multiagent system of story generation. Because this idea reflects a complex system, it is difficult to show a concrete image of the system's holistic behavior until the idea is computationally implemented. However, a simplified image of the integrative operations of various cognitive processes can be seen from the example story presented in Section 4.2.

From an integrative perspective, the following two notions will be the key points for the computational implementation of this theory:

(i) The main agent of the system is an "event" (in a broad sense) formed at various levels of abstraction, e.g., a character's action (Lisa threw away Sally's doll), a scene or a semantic segment of actions (Lisa took revenge on Sally), and the story itself.

(ii) Because the connection between agents, i.e., events or stories, is a foundation of various generative actions—

connective, contextual, gathering, and adaptive—and the integrated operations of these generative actions, the agents must be associated with each other.

6. Concluding Remarks

We proposed a new approach for the computational modeling of an autonomous intelligence based on generative narrative cognition. Throughout this paper, we conceptualized the developmental and generative process of an agent's mind as a type of self-organization on both levels: the organizational structure of a mental world and the internal structure of a story. Under this concept, as the principles of the self-organization of a mental world, we presented five types of generative actions: connective, hierarchical, contextual, gathering, and adaptive actions. Although the mechanisms of these generative actions are still abstract and implementing the proposed concept remains a distant goal, we showed the total picture of a mental system using the above concept.

The basic direction of future work will be to develop a theory of integrating generative actions in the form of a multiagent system. Agents of generative actions (in the mental system of an artificial agent) are stories and mental spaces. Although we listed many generative actions in Section 5, there are several key issues relevant to a wide array of mental activities. For example, similarity or analogy of stories is a common principle of reusing existing mental resources in various types of generative actions such as gathering actions, top-down restriction, inductive generalization, and reorganizational differentiation. Moreover, in terms of knowledge representation, how a story represents the integrated world information in an agent's mind is still unclear. The structure of a story itself, including relationships with other mental elements, needs to be considered more closely.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP18K18344 and The Telecommunications Advancement Foundation.

References

[1] B. Goertzel, R. Lian, I. Arel, H. de Garis, and S. Chen, "A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures," *Neurocomputing*, vol. 74, no. 1-3, pp. 30–49, 2010.

[2] A. V. Samsonovich, "On a roadmap for the BICA challenge," *Biologically Inspired Cognitive Architectures*, vol. 1, pp. 100–107, 2012.

[3] R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Lawrence Erlbaum, 1977.

[4] R. C. Schank, *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*, Cambridge University Press, 1982.

[5] P. Sengers, "Schizophrenia and narrative in artificial agents," in *Narrative Intelligence*, M. Mateas and P. Sengers, Eds., pp. 259–278, John Benjamins Publishing, 2003.

[6] N. Szilas, "Towards narrative-based knowledge representation in cognitive systems," in *Proceedings of 6th Workshop on Computational Models of Narrative*, pp. 133–141, 2015.

[7] C. León, "An architecture of narrative memory," *Biologically Inspired Cognitive Architectures*, vol. 16, pp. 19–33, 2016.

[8] J. de Greeff, B. Hayes, M. Gombolay et al., "Workshop on Longitudinal Human-Robot Teaming," in *Proceedings of the Companion of the 2018 ACM/IEEE International Conference*, pp. 389-390, Chicago, IL, USA, March 2018.

[9] P. Gervás, "Computational approaches to storytelling and creativity," *AI Magazine*, vol. 30, no. 3, pp. 49–62, 2009.

[10] B. Kybartas and R. Bidarra, "A survey on story generation techniques for authoring computational narratives," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 9, no. 3, pp. 239–253, 2017.

[11] M. Minsky, *The Society of Mind*, Simon & Schuster, 1986.

[12] T. Akimoto, "Narratives of an artificial agent," in *Content Generation Through Narrative Communication and Simulation*, Advances in Linguistics and Communication Studies, pp. 241–264, IGI Global, 2018.

[13] G. Prince, *A Dictionary of Narratology, Revised Edition*, University of Nebraska Press, 2003.

[14] G. Genette, *Narrative Discourse: An Essay in Method*, Cornell University Press, 1980, J. E. Lewin Trans.

[15] T. Akimoto, "Stories as mental representations of an agent's subjective world: A structural overview," *Biologically Inspired Cognitive Architectures*, vol. 25, pp. 107–112, 2018.

[16] M. Minsky, "A framework for representing knowledge," *The Psychology of Computer Vision*, pp. 211–277, 1975.

[17] P. Ricoeur, *Temps et Récit (Tome I–III)*, Seuil, 1983–1985.

[18] J. S. Bruner, *Acts of Meaning*, Harvard University Press, 1990.

[19] D. P. McAdams, *The Stories We Live by: Personal Myths and the Making of the Self*, The Guilford Press, 1993.

[20] T. Akimoto, "Computational modeling of narrative structure: A hierarchical graph model for multidimensional narrative structure," *International Journal of Computational Linguistics Research*, vol. 8, no. 3, pp. 92–108, 2017.

[21] R. R. Lang, "A declarative model for simple narratives," Narrative Intelligence: Papers from the 1999 AAAI Fall Symposium FS-99-01, 1999.

[22] S. Bringsjord and D. A. Ferrucci, *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*, Lawrence Erlbaum, 1999.

[23] N. Montfort, R. Pérez y Pérez, D. F. Harrell, and A. Campana, "Slant: A blackboard system to generate plot, figuration, and narrative discourse aspects of stories," in *Proceedings of the Fourth International Conference on Computational Creativity*, pp. 168–175, 2013.

[24] J. F. Allen, "Towards a general theory of action and time," *Artificial Intelligence*, vol. 23, no. 2, pp. 123–154, 1984.

[25] J. R. Hobbs, "Literature and Cognition," *CSLI Lecture Notes*, no. 21, 1990.

[26] W. C. Mann and S. A. Thompson, "Rhetorical Structure Theory: Toward a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.

[27] J.-D. Zucker, "A grounded theory of abstraction in artificial intelligence," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1435, pp. 1293–1309, 2003.

[28] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive Science*, vol. 7, no. 2, pp. 155–170, 1983.

[29] D. Gentner and K. D. Forbus, "Computational models of analogy," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 2, no. 3, pp. 266–276, 2011.

[30] K. J. Holyoak and P. Thagard, *Mental Leaps: Analogy in Creative Thought*, MIT Press, 1995.

[31] C. K. Riesbeck and R. C. Schank, *Inside Case-Based Reasoning*, Lawrence Erlbaum, 1989.

Hindawi
Advances in Human-Computer Interaction
Volume 2018, Article ID 7345397, 6 pages
https://doi.org/10.1155/2018/7345397

Research Article

User Experiences from L2 Children Using a Speech Learning Application: Implications for Developing Speech Training Applications for Children

Maria Uther[1,2] **Anna-Riikka Smolander**[3] **Katja Junttila**[3] **Mikko Kurimo**[4] **Reima Karhila**[4] **Seppo Enarvi**[4] **and Sari Ylinen**[3,5]

[1]*Department of Psychology, University of Winchester, Sparkford Rd., Winchester SO22 4NR, UK*
[2]*Department of Psychology, University of Wolverhampton, Faculty of Education, Health and Wellbeing, MC Building (Room MC305), University of Wolverhampton, Wolverhampton WV1 1LY, UK*
[3]*Cognitive Brain Research Unit, Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, P.O. Box 9, 00014 Helsinki, Finland*
[4]*Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, P.O. Box 13000, 00076 AALTO Espoo, Finland*
[5]*Cicero Learning, Faculty of Educational Sciences, P.O. Box 9, 00014 Helsinki, Finland*

Correspondence should be addressed to Maria Uther; m.uther@wlv.ac.uk

Received 31 May 2018; Accepted 1 October 2018; Published 1 November 2018

Guest Editor: Rafal Rzepka

We investigated user experiences from 117 Finnish children aged between 8 and 12 years in a trial of an English language learning programme that used automatic speech recognition (ASR). We used measures that encompassed both affective reactions and questions tapping into the children' sense of pedagogical utility. We also tested their perception of sound quality and compared reactions of game and nongame-based versions of the application. Results showed that children expressed higher affective ratings for the game compared to nongame version of the application. Children also expressed a preference to play with a friend compared to playing alone or playing within a group. They found that assessment of their speech is useful although they did not necessarily enjoy hearing their own voices. The results are discussed in terms of the implications for user interface (UI) design in speech learning applications for children.

1. Introduction

This study investigated child user experiences from a game-based language learning application that used automatic speech recognition (ASR) technology. The application, called 'Say it again, kid!' (SIAK) [1, 2], is designed to assist foreign language learning (vocabulary and production) in children. The technology behind the application is designed for both computers and tablets and uses ASR components for the assessment of children's speech produced while learning new words in a new, nonnative language. Thus far, the use of ASR engines in language learning has primarily been aimed at assisting language learning in native context (e.g., reading tutors such as listen [3], tball [4], space [5], and flora [6]). In itself, the use of automatic speech recognition engines in children is challenging [7, 8], but the use of ASR to aid foreign language learning is a venture that poses even further challenges and is still a field that is underdeveloped [7–11].

Nonetheless, the SIAK project sought to address this research gap by developing and implementing a foreign language learning application. The application uses a Hidden Markov Model (HMM) segmenter to find phoneme boundaries in players' utterances. Phoneme segments are individually evaluated by a Recurrent Neural Network (RNN) bilingual phoneme classifier. The classifier results are then fed into a Support Vector Machine (SVM) scoring regressor which outputs a 0-100 score, which is further mapped into a rejection or a 1-5 star score. The latency for scoring an

utterance is 2-3 seconds. The scoring mechanism is trained with in-game second-language (L2) utterances that were collected in previous experiments and scored by a human expert. The game is designed to produce a score computed using speech recognition technology after each utterance. We designed the trial of the game such that we compared performance and user reactions to a nongamified pronunciation learning environment. We used gamification since it has been suggested that this lends itself to more situated learning and a more immersive, engaging experience, which can be helpful within the language learning context (see [12–14] for a review).

The goal of the SIAK project was broadly to develop and test a novel, automatic speech recognition application to assist children in learning a new language. We have already published details of the algorithms and implementation itself [1]. In this paper, the research questions were as follows:

(i) What are the child's affective reactions to the SIAK?

(ii) What are their perceptions of pedagogical utility of gamified applications?

(iii) Where there any user aspects relating to the use of audio and interaction with speech in such applications that were problematic? This final point is crucial since the implementation of spoken language learning applications necessarily requires audio input and output.

Answering these research questions with a robust data set on user experience from actual child language learners is clearly needed in this underresearched and underdeveloped area.

As we were measuring child user experience in our application, we followed the framework for steps in usability testing according to Markopoulous & Bekker [15], namely, (1) develop assessment criteria (goals); (2) develop usability testing measurements (contexts, tasks); and (3) consider child characteristics that may constrain the design of the measurements/tasks (e.g., knowledge, age, and language ability). To this end, we focused on three areas in terms of assessment: (1) affective reactions and user engagement, (2) the perceived pedagogical value, and (3) audio interaction issues. As SIAK was an application which necessarily involved audio, we were interested in the users' perceptions of sound quality, which may then in turn affect their learning experience [16]. We were also interested in the use of speech production scoring. It should be emphasized that our goal here was not to present the actual learning outcomes (these will be presented in due course separately), but rather to report on the user experience feedback that would in turn inform the design of any future iterations of this application (and indeed speech training applications for children in general).

To measure the areas we wished to assess, we included questions regarding basic affective reactions (e.g., 'did you like this game?') as well as questions related to the elements which were pertinent to using speech-based applications (e.g., 'how clear was the sound?' and 'did you like hearing your own pronunciation during the game?'. For these questions, we used the smiley-o-meter method from the 'fun toolkit' techniques (see [17] for a review) with reference to Likert-scale answers to agree to the statement with 5 points ('not at all' to 'very much' at the extremes). There were also items from an 'again again' table (also from the 'fun toolkit'), where the player was asked questions comparing game and nongame versions of the application. The 'again again' method was chosen for comparison of the game and nongame version as it allowed freedom for the children to express a preference (or dislike) for both versions simultaneously. The 'again again' items included questions on (a) Would you like this game for yourself (yes/no or maybe)? or (b) Would your teacher like this (yes/no/maybe)? The latter question (on the teacher's presumed preferences) was selected as it has been shown to tap into the child's sense of pedagogical utility more easily than a direct question of education value. Within other studies [18], it has been shown that the perception of whether a teacher chooses an application is related to the child's perception of how good it was for learning.

In line with our research questions, we hypothesized that

(a) The user experience (in particular, affective reactions) to our new SIAK software in the target audience of 8 to 12 year olds would be positive and especially so for the game-based version of the software.

(b) Although gamification of the application would presumably add positively to the user experience, the perceived pedagogical value may be rated less positively. In addition, there may be further desire for collaborative learning as games are often enjoyed as social endeavours.

(c) Participants rating of sound quality may be affected by application or device type [16] and although not studied explicitly before to our knowledge, we also wished to explore the children's reaction to their voice being scored by computers, which may cause self-consciousness for example.

2. Materials and Methods

2.1. Participants. We recruited 117 children (59 females and 58 males) aged between 8 and 12 years (mean=9.5; SD=1.2) from the Helsinki area. The children were recruited via local schools in Helsinki, with the consent of schools to recruit and consent of the children and parents to participate. There were also some children that were recruited directly via social and community networks.

2.2. Materials. The speech learning application (SIAK) is implemented as a computer board game that runs on an Android tablet or a Windows PC (laptop) and a headset [1]. Following the testing period, the children also were given a questionnaire that modeled itself on the 'fun toolkit'—namely using the 'smiley-o-meter' scale and the 'Again again table' [17]—where they were asked to respond to a question such as 'Would you like to use this game again' responding with either 'yes', 'no', or 'maybe'. The again again items were especially used to compare game and nongame versions. There were 15 questions in total (7 smiley-o-meter items, 7 again again items and one qualitative open-ended question asking whether they

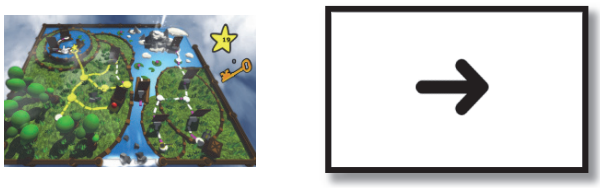


Figure 1: Screen shot of game and nongame versions of SIAK. Left panel shows example game 'world' that the learners explore. Right panel shows a nongame simple interface.

would like any particular item more (e.g., certain characters, sounds, etc.). All questionnaire items were given to the children in Finnish, which was their native language (any reference to questions in this paper are a translation into English of the original Finnish).

2.3. Procedure. The children were given the SIAK application which functioned to improve their pronunciation and broaden their vocabulary in English by introducing new English single words. As the children progressed in the game, they then encountered sentences which contained the new words. Children heard the word in Finnish and in English (produced by different native English speakers) and saw a related picture. The child was required to repeat the pronunciation of the word aloud. The children then received feedback on their pronunciation as a numerical score. The child's own and native English speaker's utterances were played again for comparison, and they received a one to five star rating based on the utterance score.

While testing, there were elements of the program that were not implemented as a game (instead of an immersive experience, they were shown a simple white background, forced order of stimulus presentation, no feedback) although the stimulation and the speech production task were the same as the game version which is described in [19]. Figure 1 shows the comparison of the screen between game and nongame version. A video of child participants playing SIAK is also available at: https://www.youtube.com/watch?v=-cgyJFV8-58&feature=youtu.be

All children evaluated the game and asked to compare game and nongame versions using the 'again again' question items (note that for 21 participants in the light user group, they did not evaluate the nongame version). The sample was divided into two groups: light users and experienced users. This was done as because we wanted a large sample of game players to give feedback on the application for user acceptance testing (UAT) and user experience (UX) reasons. For UAT/UX testing, we did not need the participants to test for many weeks at a time, but a minimum of one week – hence a 'light usage' group. On the other hand, to judge *efficacy* of the intervention, we also tested a set of 'experienced' users who have been using the application for at least 4 weeks. For this latter sample, we tested educational outcomes and brain measures as a result of training (those data will be reported separately), as training effects are typically seen over a minimum 1 month period. But it was not necessary for all participants to be tested over such a long period, which is why we had two groups. Nonetheless, the experienced user group's user experience was also tested to determine whether length of time of use might have had an impact on the user ratings.

Table 1: Affective reactions from the SIAK application using the smiley-o-meter rating scale (NB: some items had lower N due to participants not filling in every question).

Question item	N	Range of scores	Mean rating	SD
How much do you like the application? (1=not at all; 5=very much)	116	1-5	4.07	.85
How easy did you find the application? (1= very difficult; 5= very easy)	116	2-5	3.86	.78
Did you have any problems with the game? (1=very much; 5=no problems)	114	1-5	3.62	.94

For the light users, the children played approximately 10-15 min per day, 3-4 days a week. The testing period was either 1-3 weeks for light users (n=50) or 4-5 weeks for experienced users (n=67). Out of these participants, 52 used Windows laptops and 65 used Android tablets. Following the testing period, they were given the questionnaire of 15 items in their native Finnish language, which covered the breadth of three key areas: overall affective reactions; value and interest in game/nongame versions and perceived pedagogical value (indexed by question on whether their teacher would like the game for the students) and finally a set of questions relating to audio and speech interaction (e.g., perceived sound quality, utility and affective reactions to having their speech samples tested).

3. Results

3.1. Affective Reactions. From Table 1, it can be seen that the children had a generally positive reaction to the software (mean scores around 4 on a scale of 1-5, with 5 being a positive rating). There were no differences between light users and extensive users on any of these measures.

Children were also asked about affective reactions in relation to the game and nongame versions of the program using the 'Again again table'—i.e., "Would you like to use this program again?"—and they could say either 'Yes', 'No' or 'Maybe'. Data from this method was analyzed for counts in each category using chi-squared analysis and showed that there was a significant difference between the game and nongame version ($\chi 2=11.89$, $p<0.05$, $df=4$). For the game version, children were less ambivalent and generally more positive (63 out of 101 said yes to the question, 32 said maybe and only 6 said no). By contrast, for the nongame version, children were less positive and more ambivalent (only 28 out of 101 said yes, 36 said maybe and 37 said no).

Table 2: Assessment of whether teacher or self would like the game version.

	yes	maybe	no
Like game for self	59	40	13
Teacher would like the game[1]	38	72	2

[1]NB: this question was worded slightly differently in the first 50 participants but had wording changed later to ensure that the children would not misunderstand this question, from (a translation from Finnish) 'Would my teacher like this?' to 'Do you think your English teacher would like pupils to use this game'? We analysed the results from both wording versions and there was no statistically significant difference.

Table 3: Assessment of whether teacher or self would like the nongame version.

	yes	maybe	no
Like program for self	23	35	37
Teacher would like the program	22	59	14

Table 4: Rating of whether the player would prefer to play alone, with a friend or in a group.

	yes	maybe	no
Like to play alone	41	44	48
Like to play with friend	57	40	16
Like to play in a group	35	44	35

3.2. Perceived Pedagogical Utility and Collaboration Preferences for Game and Nongame Versions. The children were asked specifically about the perceived pedagogical utility by asking their views on whether their teacher would like the game versus whether they would like the game for themselves.

There was a tendency for the children to rate themselves as certainly liking the game for themselves (more 'yes' judgements) and a more ambivalent rating (more 'maybe' judgements) for the teacher ($\chi 2=22.21$, $p<0.01$), see Table 2.

For the nongame version, the children appeared to rate the self and teacher liking the game differently ($\chi 2=32.22$, $p<0.01$). In particular, they felt proportionally less ambivalent (when considering their yes/no responses) compared to the ratings the teachers who they felt would be more ambivalent and proportionally less negative, see Table 3.

Finally, when the children were asked separate questions as to whether they preferred to play alone, with a friend or with a group (results in Table 4).

It appears that in general the children appear to prefer playing with a friend more than they preferred to play alone or in a group. The differences between playing with a friend versus group might be due to the fact that 'group' may mean a group of people not known to the children and therefore they may be more ambivalent. Nonetheless, the finding that a greater proportion of positives and fewer negatives for playing with friends (compared to playing alone) would suggest that this age group tend towards preferring to play with known peers.

Table 5: Reactions from the SIAK application relating to speech and audio elements using the smiley-o-meter rating scale (NB: some items had lower N due to participants not filling in every question).

Question item	N	Range of scores	Mean rating	SD
How clear was the sound for the Finnish words? (1=not at all clear; 5=very clear)	116	2-5	4.13	0.97
How clear was the sound for the English words? (1=not at all clear; 5=very clear)	116	1-5	3.71	1.00
Did you like hearing your own pronunciation? (1=not at all; 5=very much)	114	1-5	3.83	1.21
Did you like getting feedback regarding your pronunciation? (1=not at all; 5=very much)	114	1-5	4.16	0.96

3.3. Reactions to Elements Related to Speech Training Software. In this category, there were two sets of questions posed to the children that are useful to consider when designing speech training software. The first set was in relation to the rating of speech quality. Here, the children were asked two separate questions: 'How clear was the sound for the Finnish-speaking words?' and 'How clear was the sound for the English-speaking words?'. The second set of questions related to the use of speech pronunciation feedback. As the program involved not only getting scoring as feedback, but also a replay in comparison to the native speech, they were asked two questions: 'Did you like getting feedback regarding your own pronunciation?' and 'Did you like hearing your own pronunciation during the game?'

Interestingly, the children rated the Finnish samples as having better sound quality than the native English samples ($F_{1,112}=7.703$, $p<0.01$), see Table 5. Although the sampling rate, microphone frequency responses were the same, they were not collected in identical labs and hence there may have been subtle differences. However, it appears that experience with the samples may have also played a part in the rating. The more experienced users rated samples better compared to the less experienced users of the application ($F_{1,112}=4.739$, $p<0.05$). This would suggest that the tendency to rate English sounds as worse quality than the Finnish ones might be due to exposure to that particular language.

With respect to the other questions regarding pronunciation, we looked at the aspects of perceived helpfulness of pronunciation training versus the actual experience of hearing their own voice. As one might have predicted, the children liked getting feedback more than the process of actually hearing their own voice. This is probably due to the children feeling self-conscious about their voices, but yet seeing the value of feedback.

4. Discussion

The results from our extensive user trial show that children in general had a positive user experience from the speech training SIAK game, answering our first question regarding the affective responses of children in this age group to the application. There was evidence that they in general were positive about the application (scoring over 4 out of 5 on a scale of 1-5) and they had found the application fun and helpful. They also did not report major difficulties with the application and found it easy to use. This is encouraging as it is helpful to have a tool that is perceived to be easy to use and elicits positive user feedback in this group.

With respect to our second research objective, the game version was perceived more positively in this age group compared to the nongame version. Interestingly though, the children ranked the perceived pedagogical value (marked by the question of whether the teacher liked the game) as being more definitely more negative to the nongame version than the game version. This contradicts our initial hypothesis that the game-based version may be perceived as having less 'educational value' and therefore be perceived by the children as being less favourably rated by their teachers. On the other hand, this was coupled with the finding that the children rated their teachers as more ambivalent towards the game (and nongame) versions than the children, suggesting that children may not necessarily be clear as to what their teachers would think. When interpreting these results, we need to be mindful that the children thought the question was meant to ask whether the teachers would choose the application for themselves, whereas instead, the intended focus was to ask whether the teachers would choose the application for the child. As [20] states, usability testing in children can often lead to unexpected results and raise more issues in testing than originally envisaged. However, this explanation is unlikely given that an elaboration of the wording was made in a later stage of data collection to clarify understanding, which did not change the results. Of course, further research is needed to definitively tell whether the children understood the question in the way intended. For example, one possibility could be that 'theory of mind' might not be fully operational in the children at the lower end of the age range of the sample [21]. However, we did not see any age differences in the way these questions were answered either.

With respect to the issue of collaboration, it appeared that this age group slightly preferred interaction with friends. Although more data would be needed to confirm this, the trend seen in these data accord with the results of other studies (e.g., Heikkinen et al.'s. JamMo implementation [22]) which showed that children of this age group report liking working in pairs or very small groups, particularly on mobile devices.

With respect to the third research question, we also sought to investigate whether there were any auditory interaction issues (perceptions of sound quality, experience of pronunciation feedback) that may impact on the user learning experience or learning outcomes. With respect to the listening experience: it appears that there were some perceptions of difference in sound quality between native and nonnative speech. Although at the time of writing, detailed acoustical analysis on the speech was not available to know whether the perceived differences were the result of actual real subtle differences, the effect seemed to be moderated by experience (in other words, the heavy users of the program rated the nonnative speech better in sound quality than the more naïve users of the program). That is, the perception of sound quality may be affected by the difficulty to map speech input into existing mental representations. Further research is needed to determine whether sound quality is being affected by language experience.

With respect to pronunciation feedback, it was clear that this age group found feedback useful, but was less positive about hearing their own voice. These less positive ratings could be potentially mitigated by including assurances for the learners about confidentiality and the value of receiving a replay of their own voice. It is unclear whether this might reflect performance anxiety which may in turn see effects on performance. Further analysis on actual learning outcomes could explore whether there is a relationship as has been seen in other contexts [23].

In summary, future research will need to focus on the following areas:

(1) Whether user perceptions of voice clarity occur in native versus nonnative language in other samples and differ as a function of nonnative language experience (as we found here). We would also be interested in investigating whether such biases in perceived clarity impacts negatively on learning outcomes in foreign language learning contexts.

(2) Whether children in general are self-conscious of automated recognition as we found here and it would be useful to know whether such effects are modulated by age.

(3) Whether the positive affect ratings for gaming result in improved language learning outcomes compared to nongamified versions of the implementation. Such data would be helpful to determine whether the value of gaming in other domains also transfers to automatic speech recognition.

(4) Finally, we would be also interested in exploring whether children in general prefer to work with their friends in all learning contexts – or are there some situations where they may prefer to work alone (e.g., when they are being assessed and could be self-conscious).

5. Conclusions

In conclusion, these data serve a starting point of observations around speech training elements that are useful and well received for this 8-12-age group. Further work around the perception of speech quality in nonnative language and preference for collaboration is needed. What is clearly confirmed from the data is that the gaming aspect of the application is well-received and serves well as a positive tool to deliver speech training in this group. Children also interestingly

appear to rate their teachers as being more ambivalent but also less negative for the game version. This may suggest that they do not necessarily perceive the game aspect to have less pedagogical value than the nongame version, although further research is needed to clarify.

Data Availability

The anonymised, de-identified, raw data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

A subset of preliminary data from this study were disseminated in an oral presentation and abstract form at Eurocall 2018, August 2018, Jyväskylä, Finland.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors gratefully acknowledge funding from Academy of Finland (Grant no. 1274058) and the funding from Visiting Professor Grant awarded to Professor Uther by Nokia Foundation to complete the work on this evaluation while on a visit hosted by Cicero Learning.

References

[1] K. Reima et al., "SIAK - A Game for Foreign Language Pronunciation Learning," *Interspeech*, vol. 2017, 2017.

[2] A. Rouhe, R. Karhila, K. Heini, and M. Kurimo, "A pipeline for automatic assessment of foreign language pronunciation," *ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, vol. 2017, 2017.

[3] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "Prototype reading coach that listens," in *Proceedings of the 12th National Conference on Artificial Intelligence. Part 1 (of 2)*, pp. 785–792, August 1994.

[4] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association, Interspeech 2007*, pp. 941–944, Belgium, August 2007.

[5] J. Duchateau, Y. O. Kong, L. Cleuren et al., "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Communication*, vol. 51, no. 10, pp. 985–994, 2009.

[6] D. Bolan-Os, R. A. Cole, W. Ward, E. Borts, and E. Svirsky, "FLORA: Fluent oral reading assessment of children's speech," *ACM Transactions on Speech and Language Processing*, vol. 7, no. 4, 2011.

[7] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction, WOCCI '09*, USA, November 2009.

[8] J. Proença et al., "Mispronunciation Detection in Childrens Reading of Sentences," *IEEE/ACM Trans. AUDIO*, vol. 26, no. 7, pp. 1203–1215, 2018.

[9] E. M. Golonka, A. R. Bowles, V. M. Frank, D. L. Richardson, and S. Freynik, "Technologies for foreign language learning: A review of technology types and their effectiveness," *Computer Assisted Language Learning*, vol. 27, no. 1, pp. 70–105, 2014.

[10] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech and Language*, vol. 23, no. 1, pp. 65–88, 2009.

[11] S. M. Abdou, S. E. Hamid, M. Rashwan et al., "Computer aided pronunciation learning system using speech recognition techniques," in *Proceedings of the INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, pp. 849–852, USA, September 2006.

[12] F. Cornillie, S. L. Thorne, and P. Desmet, "ReCALL special issue: Digital games for language learning: Challenges and opportunities," *ReCALL*, vol. 24, no. 3, pp. 243–256, 2012.

[13] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers & Education*, vol. 59, no. 2, pp. 661–686, 2012.

[14] B. DaCosta and S. Seok, "Developing a Clearer Understanding of Genre and Mobile Gameplay," in *Handbook of Research on Immersive Digital Games in Educational Environments*, Advances in Educational Technologies and Instructional Design, pp. 201–231, IGI Global, 2019.

[15] P. Markopoulos and M. Bekker, "On the assessment of usability testing methods for children," *Interacting with Computers*, vol. 15, no. 2, pp. 227–243, 2003.

[16] M. Uther and A. P. Banks, "The influence of affordances on user preferences for multimedia language learning applications," *Behaviour & Information Technology*, vol. 35, no. 4, pp. 277–289, 2016.

[17] G. Sim and M. Horton, "Investigating children's opinions of games," in *Proceedings of the the 11th International Conference*, p. 70, Bremen, Germany, June 2012.

[18] G. Sim, S. MacFarlane, and J. Read, "All work and no play: Measuring fun, usability, and learning in software for children," *Computers & Education*, vol. 46, no. 3, pp. 235–248, 2006.

[19] R. Karhila, "SIAK - A game for foreign language pronunciation learning," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017, p. 2017, 2017.

[20] G. Sim, J. Read, and M. Horton, "Practical and ethical concerns in usability testing with children," in *Games User Research: Study Approach, Case*, A. Peters, Ed., pp. 1–33, CRC Press, 2016.

[21] C. I. Calero, A. Salles, M. Semelman, and M. Sigman, "Age and gender dependent development of theory of mind in 6- to 8-years old children," *Frontiers in Human Neuroscience*, no. MAY, 2013.

[22] K. Heikkinen, J. Porras, J. Read, and G. F. Welch, "Designing Mobile Applications for children," in *User Requirements for Wireless, No. September*, L. T. Sorensen and and K. E. Skouby, Eds., pp. 501–512, River publishers, Aalborg, Denmark, 2015.

[23] D. E. Callan and N. Schweighofer, "Positive and negative modulation of word learning by reward anticipation," *Human Brain Mapping*, vol. 29, no. 2, pp. 237–249, 2008.

Hindawi
Advances in Human-Computer Interaction
Volume 2018, Article ID 8406187, 10 pages
https://doi.org/10.1155/2018/8406187

Research Article

Student Evaluations of a (Rude) Spoken Dialogue System Insights from an Experimental Study

Regina Jucks, Gesa A. Linnemann, and Benjamin Brummernhenrich

University of Muenster, Institute of Psychology for Education, Germany

Correspondence should be addressed to Regina Jucks; jucks@uni-muenster.de

Received 21 January 2018; Revised 1 June 2018; Accepted 11 June 2018; Published 1 August 2018

Academic Editor: Thomas Mandl

Communicating with spoken dialogue systems (SDS) such as Apple's Siri® and Google's Now is becoming more and more common. We report a study that manipulates an SDS's word use with regard to politeness. In an experiment, 58 young adults evaluated the spoken messages of our self-developed SDS as it replied to typical questions posed by university freshmen. The answers were either formulated politely or rudely. Dependent measures were both holistic measures of how students perceived the SDS as well as detailed evaluations of each single answer. Results show that participants not only evaluated the content of rude answers as being less appropriate and less pleasant than the polite answers, but also evaluated the rude system as less accurate. Lack of politeness also impacted aspects of the perceived trustworthiness of the SDS. We conclude that users of SDS expect such systems to be polite, and we then discuss some practical implications for designing SDS.

1. Introduction

Advances in speech recognition move our interactions with spoken dialogue systems (SDS), such as Apple's Siri or Google Now, ever closer to human dialogue. Over and above the capability of conversing in natural language, one strain of development concerns genuinely social aspects of human communication, such as alignment, addressing by name, and politeness. This paper addresses how students evaluate the communication behavior of an SDS that employs the social strategies of politeness and rudeness. Their evaluations address variables such as acceptance, appropriateness, and competence, all of which are also relevant in evaluating human behavior. In the following, we briefly address the technological aspects of SDS. Then we outline politeness theory and introduce how speakers take their speech partners' autonomy and affiliation into account when formulating messages. Finally, we provide results of our empirical research on the perception of SDS, which offer insights about how to design them effectively.

1.1. Technology That Interacts: Insights into the Mechanisms and Usage of Spoken Dialogue Systems. More and more computers are now able to communicate with their users in natural language [1]. For example, spoken dialogue systems (SDS) serve as personal assistants and are implemented in smartphones (like Siri from Apple and Cortana® from Microsoft), cars [2], or special devices (such as Echo from Amazon: [3]). SDS are also used in educational contexts, e.g., in learning programs for children [4].

The first attempts at emulating human communication were simple chatbots such as Eliza which tried to emulate a psychotherapist in the Rogerian, person-centered tradition [5]. How chatbots can be made more human-like is still a very active field of development, for example, by evaluating the conditions of *uncanny valley* reactions in animated and text-based bots (e.g., [6]). Since 1990, the Loebner Prize has been held as an annual competition with an award given to the creator of the chatbot that most convincingly acts as a human interlocutor.

Whereas chatbots are usually text-based and often have entertainment purposes, SDS serve as an interface to a specific system and let computer systems and human users interact using spoken language; that is, the computer systems are capable of understanding and producing spoken language. This can reach from rather simple "command-and-control"

interactions, which refer to short, single controlling functions (e.g., [4]), to systems that are able to handle complex input and to generate complex, natural language (e.g., [7, 8]).

SDS have become much better at mimicking human interactions. Incrementality allows the system to implement backchannels and barge-ins. These are associated with facilitated *grounding* [9] in human-human interaction, meaning that a user is more confident about whether the system shares the user's understanding off the topic of the conversation and the meaning of the words used in it [7]. Connecting and operating SDS via the Internet can help support the system and the user in a concrete situation, e.g., by searching the Internet for the answer the SDS does not "know" directly. Additionally, such searches can be analyzed and used to improve the SDS [1]. Due to their proactive behavior, SDS are no longer restricted to simple replies. They can now initiate conversations themselves and provide information without being asked [1]. Thus, concerning their language, SDS can possess a high degree of anthropomorphism [10–12].

Providing these characteristics, SDS can be employed as tutors, termed intelligent tutoring systems. Graesser and colleagues [13] recently introduced an intelligent tutoring system from a new, "mixed-initiative" generation. Their conversational agent is able to maintain interactions over multiple turns, presenting problems and questions to the learner. Thus, it promotes active knowledge construction and outperforms mere information-delivery systems. Besides its conversational capabilities, social aspects of SDS play an important role. Research on intelligent tutoring systems has recently turned its attention on agents that can act socially and respond to tutees' affective states (e.g., [14, 15]). When intelligent tutors and tutees are teammates, for instance, tutoring seems to be more effective [16]. Put differently, how intelligent tutors—and SDS in general—communicate plays an important role. Politeness has seemingly been neglected in conversational agents [17]. However, some computer tutors already exhibit different kinds of polite instruction [18]. In the next section, we will give an overview of politeness theory and its effects on communication.

1.2. Mitigating Face Threats: Insights into Politeness Theory. Communication involves more than just conveying information. It is a social activity that contributes to individuals' social needs and impacts their self-concepts. In their politeness theory, Brown and Levinson [19] argue that every person has a public self-perception, the so-called *face*, and that this face needs to be shaped positively. Individuals' needs for belonging and support affect the *positive face*, and the need for autonomy and freedom of action affect the *negative face*. Communication behavior that serves as support of a person's face is termed *face work*.

Actions that harm and affect someone's face are called *face-threatening acts* (FTAs). To what extent an FTA is perceived as serious depends on social distance, power relations, and the absolute imposition. During communication, every single contribution might include FTAs, be it through direct orders, which restrict autonomy and independence, or through corrections, which in a way serve as a rejection of the person. Politeness theory describes several strategies for how to mitigate an FTA [20]. The first option is not to perform the FTA and, accordingly, the utterance (the topic/aspect is simply not addressed). Second, the FTA can be transmitted *off record*; that is, the speaker remains vague or ambiguous. Hence he/she cannot be sure if the intended meaning is conveyed. Third, negative politeness is used, which can reduce the direct imposition on the hearer (e.g., through apologizing, employing hedges, and being conventionally indirect). This includes formulating a request indirectly as a suggestion. Fourth, positive politeness aims to meet the hearer's need for belonging and at gratification. A message might be started with positive feedback on the communication partners' intelligence. Mitigating face threats whenever autonomy or affiliation is threatened is a natural communication behavior.

1.3. Human-Like Interaction? Empirical Research on the Perception of SDS. As SDS possess enormous capabilities, users are likely to consider them conversational partners. SDS are perceived with their human-like qualities even when users are aware that they are communicating with a computer [21]. Basic language capabilities, for example, giving simple responses such as "yes" and "no", seem to be enough to perceive the computer as a human-like being [22]. When users perceive the SDS as a competent partner, they also perceive it as a social actor [23]. In this way, an SDS with a female voice can, for example, potentially activate corresponding gender stereotypes [10]. Thus, it becomes relevant how users assess an SDS according to different aspects that humans are evaluated on. One aspect that also directly influences how the information conveyed by the SDS is processed is *trustworthiness*, which includes **a**bility, **b**enevolence, and **i**ntegrity (see [24] and the ABI model; [25]).

If computers are perceived as social agents, this influences our interaction with them. For example, people prefer systems that communicate in a personal manner [21]. People also usually communicate politely with computers and avoid explicit face threats (e.g., [26]) but also sometimes lie or behave intentionally rudely toward them [27]. This could be caused by disinhibition but could also be a reflection of the prevalence of rudeness in human communication [28].

Given that people tend to treat computers as humans, are social behaviors such as politeness also expected from SDS [29]? In human communication, politeness improves social perceptions. Polite speakers appear more likable and recipient oriented [30, 31]. Some studies have also shown effects on attributes such as perceived integrity or competence [32]. Thus, if an SDS is perceived similarly to a human interlocutor, similar effects should be found.

Regarding trust, it can be argued that SDS can be conceptualized as receivers of trust, as trustees (see also [33]). McKnight [34] has stated that "trust in technology is built the same way as trust in people" (p. 330). In the vast majority of conceptualizations, trust incorporates the willingness to be vulnerable [35] and therefore the willingness to depend on somebody else (e.g., [36]).

Regarding politeness, users tend to communicate politely with computer systems (e.g., [26]). Pickard, Burgoon, and Derrick [37] have found that likeability, in turn, increases the

tendency to align, e.g., using the same words and expressions to a conversational agent. Lexical alignment is perceived as polite [38, 39] and evokes positive feelings [38, 40]. Gupta, Swati, Walker, and Romano [41] developed a system which employs artificial spoken language and politeness principles in task-oriented dialogues. They found, for example, that the predictions of politeness theory are applicable to discourse with such systems: the strategies had an impact on the perception of politeness (see also [42]). However, we are unaware of any investigations into intentionally rude systems, with the exception of a rude intelligent tutoring system that has proven beneficial for some, but not all, of its students [43].

Nowadays, companies offer the possibility to personalize SDS (e.g., Siri can be taught the user's name and relationships; navigation systems' language style can be adjusted, e.g., restricted and elaborated language style: NIK-VWZ01 Navigation, n.d.). De Jong, Theune, and Hofs [44] developed an embodied conversational agent that is able to align to the politeness shown by its interlocutor. Some participants appreciated that the system aligned in terms of politeness, others preferred a version of the system that always displayed the same high degree of politeness, irrespective of whether the user showed no politeness.

2. Rationale

The above-reported literature strongly shows that communication with SDS is part of everyday life. Politeness theory has introduced the concept of face threats, e.g., affronts to a person's autonomy and/or needs of belonging. These face threats are mitigated via communicational behavior, such as hedges [45] and relativizing words.

Empirical studies indicate that SDS are evaluated on social dimensions comparably to humans. Those experimental studies have also shown that word usage impacts this evaluation.

The following study manipulates SDS word usage with regard to politeness. We conducted a 1x2 experimental design, where politeness was contrasted with rudeness. We define rudeness as a deliberate attack on the addressee's face that can be used both playfully and aggressively [46]. In this respect, rudeness is distinct from a mere lack of politeness when uttering a necessary face threat (i.e., a bald/on record strategy in the terms of politeness theory), as well as an unintentional face threat (e.g., an SDS asking a strongly religious user a question about a topic that is offensive to them) but consists of intentionally aggravating behavior. The interpretation often is dependent on the context (e.g., [47]).

We formulate three hypotheses that address participants' evaluation of both the social aspects of the SDS and its competence. Furthermore, by directly asking participants to suggest changes in SDS formulations, a direct measure on the word level was operationalized.

H1: A polite SDS will be judged as more likable and polite than will a rude SDS and its responses as more appropriate and pleasant than those from a rude SDS. A polite SDS will be more strongly perceived as a social interaction partner than a rude SDS.

H2: A polite SDS will be judged as more competent and trustworthy than a rude SDS.

H3: Rude responses will be perceived as more serious face threats than polite responses and will lead to more revisions than polite ones.

In the following, methods and materials are described.

3. Methods and Materials

In order to comply with the requirements of open science and to achieve transparency, we report how we determined our sample size, all (if any) data exclusions, all manipulations, and all measures in the study [48].

3.1. Participants. We recruited participants at an open house event that our university holds yearly for potential new students. We planned to recruit as many participants as possible, aiming for about 80. This would yield a power of about 80% in finding medium to large effects, as previous studies on similar phenomena have found.

In total, 58 persons participated in the study (35 females). The age of the participants ranged from 15 to 20 years old and the mean age was 16.91 (SD = 1.08) years old. All were native German speakers or spoke German since their early childhood. In Germany, students that aim for university entrance qualifications usually choose two intensive courses during the last two years of school. In our sample the most common choices were biology (36% of participants), English (33%), German (29%), and mathematics (21%). They reported to use a computer on average of 9.84 hours per week (SD = 9.42) and to use the Internet on average of 25.34 hours per week (SD = 21.17). Also, 86% of participants considered themselves to possess intermediate or advanced computer knowledge. Of all participants, 12% reported that they use SDS on a daily basis or several times a week, 8.6% reported to use SDS several times a month, and 77.6% reported that they rarely or never use SDS. Overall, 66% of participants indicated that they use SDS that are implemented in smartphones, like Siri and Google Now.

Experimental conditions did not differ for each of the above-reported descriptive variables, all $Fs > 0.832$, $ps > .366$. Hence, these variables were not considered further on. No data were removed from the analysis. Participants received no compensation.

3.2. Materials. The study was conducted at a German university; all materials were in German. Materials and stimuli are available at https://sites.google.com/site/sdspoliteness. The examples we present in this paper are translations of the original materials. The speech was created using a text-to-speech synthesis tool available on Apple Mac computers. It was a female voice.

We told participants that we built/designed and trained an SDS to provide answers to university freshmen's questions. We stated that we therefore wanted them to assess the SDS's answers. The participants were told that the SDS was part of a mobile app provided to new psychology students to help them find their way around the university and their courses. To this end, users could ostensibly ask the SDS questions about

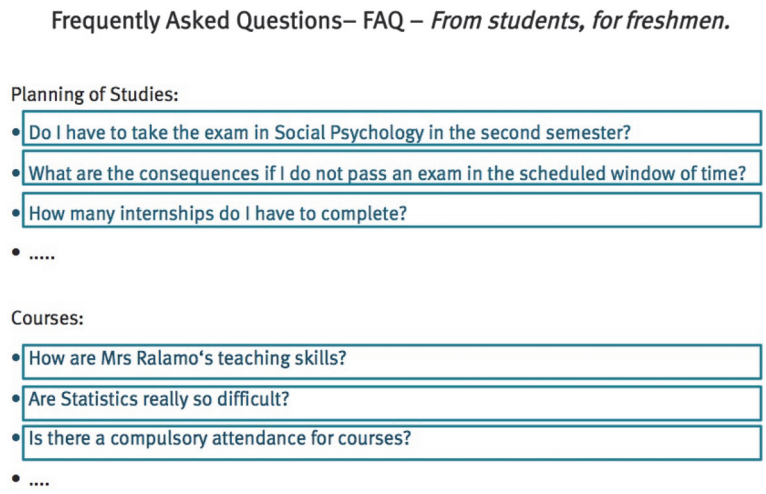


Figure 1: Screenshot of the question list shown to participants.

different topics related to university life and the psychology program and receive informative answers. In reality, the app does not exist and participants were debriefed at the end of the experiment.

Participants listened to six answers presented by our own designed SDS, called ACURI. The answers addressed six typical questions of university freshmen. The questions were presented on a screen and after clicking on the respective question, the answer was presented as a sound file. The questions were the same in both experimental conditions, but the answers given by the SDS differed. In the *polite condition* the messages were phrased politely (e.g., "You can choose whether you want to..."). In the *rude condition* the face threats were strong(er), resulting in a rudely phrased message (e.g., "No, you have to do. . ."). Questions and answers as well as the original versions in German are shown in Table 1; see Table 2 for an example.

3.3. Procedure. Participants were tested in groups of five or six. Each member of the group was assigned the same experimental condition. A 1x2 between-subject design was realized, with $n = 29$ participants in each of the two conditions (polite versus rude). All participants were seated individually in front of a laptop computer and were provided with headphones. They received sheets containing information on the experimental procedure as well as a declaration of consent and data privacy. They also received a booklet with the questionnaire to be completed as part of the study. Our local ethics committee approved the study.

After reading the information and signing the form, the students' questions were presented on the computer screen. The participants were instructed to click on the questions to hear the SDS's responses. The responses were played as audio over the headphones. Participants were not free to choose which question to click; there was a fixed order in which the answers to the questions (sound files) were available (see Figure 1 for an example of a participants' screen).

After every response from the SDS, the participants were asked to turn a page of the booklet and respond to a number of questions concerning the response they just heard. After all responses had been heard, the participants were asked to rate the SDS per se and provide demographic information. After completing the questionnaire, the participants were debriefed. The session took about 30 minutes.

3.4. Dependent Measures. Participants were asked to rate every single response of ACURI on the perceived face threat as well as the holistic impression of ACURI at the end of the survey.

3.4.1. Evaluation of the Responses. The participants rated each of the six responses on the following.

Pleasantness and Appropriateness. Participants were asked to indicate on 7-point bipolar semantic differentials whether they found the response (1) pleasant–unpleasant and (2) appropriate–inappropriate.

Perceived Face Threat Scale (PFT; [49]). The scale was originally designed to rate utterances in workplace conversations and later complaints in interpersonal contexts, regarding how much they threatened positive and negative face aspects. The authors found that responses on the scale differed depending on the type of the complaint, such that complaints that had focused on the disposition of the receiver were judged as more face-threatening. This makes the scale adequate for our goals. We are unaware of any other measure for evaluating face threat.

Sample items are "My partner's actions showed disrespect toward me" for threatening positive face and "My partner's actions constrained my choices" for threatening negative face. Because we did not pose a hypothesis regarding the different face aspects, we averaged scores on all 14 items into a single value. For the current study, all items were translated into German and adapted to collect ratings on "ACURI" instead of "my partner". Participants responded on a 7-step Likert scale ranging from "not at all" to "very much". Scale reliability was good with Cronbach's $\alpha = .91$.

Table 1: Student's questions and polite and rude phrasings of face threats in ACURI's responses.

Student Question	Polite Phrasing in ACURI's Response	Rude Phrasing in ACURI's RESPONSE
Do I have to take the exam in Social Psychology in the second semester?	You can choose whether you want to do the exam in your second or third semester. You find this information on the website if you look for the link to syllabus and exam information.	No, you have to do the social psychology exam in the second or third semester. You find this information on the website if you look for the link to syllabus and exam information.
What are the consequences if I do not pass an exam in the scheduled window of time?	That can happen but it's regrettable. Other courses can only be taken once the exam is passed.	Well, hard luck! You have to pass the exam before you can take other courses.
How many internships do I have to complete?	For your Bachelor's degree this is around 12 weeks. The exact number of hours varies from Bachelor to Master degree. If you get a chance take a look at the syllabus. After completing your internships you will for sure have a good idea about the different fields Psychologists work in.	As many as you need in order to get an idea of the different fields Psychologists work in. The exact number of hours varies from Bachelor to Master degree. To find out you will have to look into the syllabus.
How are Mrs. Ralamo's teaching skills?	Students she perceives to be motivated to work will pass her class rather easily. So you will get a long with her well if you do the work in class.	Well it's going to be difficult for lazy ones. Make an effort and do the work in class and you won't have problems with her.
Are Statistics really that difficult?	If you make an effort you will do okay. In addition to that there is a statistics tutorial that you can enroll in.	People asking such a question will probably have problems. If you are already worried that you won't manage it'll probably turn out this way. Enroll in the statistics tutorial right away; the tutors regularly make miracles happen.
Is there a compulsory attendance for courses	For most lectures and classes attendance is not mandatory but you should try to go nonetheless because contents to be learned usually stick better this way then through self-study.	No, attendance is not mandatory for most lectures and classes. But how do you expect to learn if you don't go to class?

Table 2: Example of a student's question (SQ) and the polite and rude answers, respectively, of the spoken dialogue system named ACURI.

	Polite	Rude
SQ1	Do I need to take the social psychology exam in the second semester?	
ACURI	You can choose whether you want to do the exam in your second or third semester. . . .	No, you have to do the social psychology exam in the second or third semester. . . .

3.4.2. Evaluation of the SDS. The participants evaluated the SDS regarding several subjective appraisals, epistemic trust, and whether they perceived it as a social agent.

Subjective Assessment of Speech System Interfaces. The participants rated the SDS on the Subjective Assessment of Speech System Interfaces measure (SASSI, [50]). This instrument was developed as an extension of earlier measures to evaluate the usability of graphic interfaces and focuses on subjective aspects of SDS usability. It consists of six subscales: response accuracy (nine items, e.g., "the system makes few errors"), likability (nine items, e.g., "the system is pleasant"), cognitive demand (five items, e.g., "a high level of concentration is required when using the system"), annoyance (five items, e.g., "the interaction with the system is frustrating"), habitability (four items, e.g., "I was not always sure what the system was doing"), and speed (two items, e.g., "the interaction with the system is fast"). There are alternative measures focusing on SDS usability, most notably the CCIR-BT [51] with similar subscales. However, the SASSI measure has been more widely used in subsequent research (e.g., [52]).

In the current study, the participants indicated their agreement to the statements on 7-point Likert scales. Scale reliabilities for the response accuracy and likability subscales were good, with α = .84 and .87, respectively. Reliabilities for the cognitive demand and annoyance subscales were acceptable with α = .68 and .72, respectively. However, the subscale reliabilities for the habitability and speed subscales were inadequate with α = −.06 and .12, respectively. This might be explained by the fact that the participants were not using the SDS themselves but merely judging the SDS's utterances. The two subscales were dropped from further analyses.

Hence, four measures, each one for a subscale of SASSI serves as subjective assessment of ACURI.

Perceiving the SDS as a Social Agent. The participants indicated how much they perceived the SDS as a social agent using a measure developed by Holtgraves, Ross, Weywadt, and Han [21]. This inventory was developed in order to assess whether users ascribe human-like qualities to chatbots. The measure consists of two subscales that measure perceptions of conversational skill (three items, e.g., "how engaging is the system?") and pleasantness (three items, e.g., "how thoughtful is the system?"). The participants responded on 7-point semantic differentials (e.g., "not at all thoughtful–very thoughtful"). In the original publication, chatbots using the user's first name were found to be evaluated more positively on these scales. In our study, subscale reliabilities were good to satisfactory with Cronbach's α = .67 and .81, respectively.

Epistemic Trust. The participants rated how much they trusted the SDS as a source of knowledge using the Münster Epistemic Trustworthiness Inventory (METI; [53]). This measure is based upon the ABI model mentioned in the introduction [24] and consists of 5-point bipolar adjective pairs. The subscales measure goodwill (four items, e.g., "moral–immoral"), expertise (six items, e.g., "qualified–unqualified"), and integrity (five items, e.g., "honest–dishonest"). All subscales exhibited satisfactory consistencies with Cronbach's α =.73, .81, and .60, respectively.

There are alternative measures that measure similar constructs, such as the Credibility scales by McCroskey and Teven [54]. However, the METI instrument differs from these in that it explicitly focuses on *epistemic* trust, that is, whether the target of the evaluations is a trustworthy source for the knowledge that the user seeks. This was desirable for our research question.

Suggestions for Rephrasing. Participants were given the opportunity to rephrase ACURIs answers and to mention things that, from their perspective, should be changed. They provided their answers in writing as response to this instruction: "Please listen to the answer of the question once again. You may now note potential suggestions for changing the answer."

For the analysis, we used a bottom-up, data-driven process to identify five categories of statements of what should be changed according to the participants: (1) apply strategies to mitigate FTAs (e.g., "use 'it is recommended that you...' instead of 'you have to...'"); (2) formulate the utterance in a more direct and neutral way, e.g., "should just answer the question and not make a proposal"; (3) change the prosody or pronunciation, e.g., "a brighter and less monotonous voice would be better"; (4) provide more precise information, e.g., "give information about where the tutorial takes place"; and (5) no changes are necessary, e.g., "the hint was helpful".

4. Results

We used the lme4 package [55] for the R statistical software and entered the six individual responses as a random effect and the politeness condition (polite/rude) as a fixed effect into linear mixed-effects models. For the ratings collected after the whole discourse, we calculated ANOVAs and MANOVAs, depending on the measure as reported below.

With ***hypothesis 1***, we assumed that a polite SDS would be judged as more likable and polite responses as more appropriate and pleasant than a rude SDS. Furthermore, we expected a polite SDS to be more strongly perceived as a social interaction partner than a rude SDS. These expectations

were partly confirmed: both ratings of pleasantness and appropriateness measures yielded the expected effects. Polite responses were perceived as more appropriate ($M = 6.53$, $SE = 0.28$) than rude responses ($M = 5.59$, $SE = 0.28$), $F(1,55)=13.59$, $p < .001$. Polite responses were also perceived as more pleasant ($M = 6.07$, $SE = 0.28$) than rude responses ($M = 4.97$, $SE = 0.29$), $F(1,55) = 10.43$, $p < .001$.

Contrary to our expectations, on the holistic level, the SASSI likability subscale did not show a significant effect for the politeness condition, $F(1,51) = 1.87$, $p = .177$. Both groups judged likability of ACURI as moderate (polite: $M = 4.62$, $SD = 1.07$; rude: $M = 4.18$, $SD = 1.19$).

Regarding how much participants perceived the SDS as a social agent with human-like properties, we used the measure by Holtgraves and colleagues [21]. The pleasantness subscale showed a significant effect of the politeness condition in the expected direction. As such, the polite SDS was also judged as more pleasant ($M = 5.67$, $SE = 0.23$) than the rude SDS ($M = 4.14$, $SE = 0.23$), $F(1,51) = 17.88$, $p < .001$. However, the polite SDS's conversational skill was not judged to be higher, $F(1,51) = 1.04$, $p = .31$.

With ***hypothesis 2***, we assumed that a polite SDS would be judged as more competent and trustworthy than a rude SDS. This hypothesis was mostly confirmed. The polite SDS received higher ratings on the SASSI response accuracy subscale ($M = 4.65$, $SE = 0.13$) than the rude SDS ($M = 4.23$, $SE = 0.13$), $F(1,51) = 4.48$, $p = .039$. The MANOVA with the METI subscales showed a significant effect for the politeness condition, $F(3,49) = 4.15$, $p = .011$. The follow-up analyses showed that the polite SDS was judged as showing more goodwill (polite: $M = 3.63$, $SD = 0.72$; rude: $M = 2.90$, $SD = 0.78$) and integrity (polite: $M = 3.90$, $SD = 0.63$; rude: $M = 3.51$, $SD = 0.58$) but not more expertise (polite: $M = 3.70$, $SD = 0.71$; rude: $M = 3.54$, $SD = 0.73$) than the rude SDS.

With the ***hypothesis 3***, we expected that rude responses would be perceived as more serious face threats than polite responses. H3 was confirmed. According to the PFT, rude responses were perceived as more face-threatening ($M = 3.22$, $SE = 0.14$) than polite responses ($M = 2.46$, $SE = 0.14$), $F(1,55) = 22.46$, $p < .001$.

The average amount of revision proposals regarding all respective answers was comparable in both conditions. On average, participants made suggestions on 2.48 ($SD = 2.13$) answers in the polite condition and on 2.62 answers in the rude condition ($SD = 1.97$; $F(1,56) = 0.07$, $p = .799$). However, the amount of suggestions regarding the categories differed depending on the respective response (see Table 3 for details). We calculated Fisher's exact tests for each response to test for differences between conditions. This is a statistical technique for the analysis of deviations from expectation in a contingency table [56] and is especially suited for smaller samples.

The condition had an influence on answers to SQ 2 ($\chi^2(4, N = 58) = 16.550$, $p < .001$), SQ4 ($\chi^2(4, N = 58) = 9.184$, $p = .029$), and SQ5 ($\chi^2(4, N = 58) = 12.025$, $p = .006$), but not on answers to SQ1 (SQ1: $\chi^2(4, N = 58) = 2.545$, $p = .725$, *ns.*), SQ3 ($\chi^2(5, N = 58) = 8.086$, $p = .110$, *ns.*), and SQ6 ($\chi^2(4, N = 58) = 5.463$, $p = .213$ *ns.*). In the rude condition, most changes were aimed at strategies to mitigate FTAs (29 % on average). In the politeness condition, most suggestions were offered in categories 3 or 4. Participants referred to providing more precise information and on changing or enhancing the employed voice (both categories each around $M = 14\%$). These two categories were each chosen for 6% of the answers by participants of the rude condition.

5. Discussion

To sum up, the results of this study show that polite responses were perceived as more appropriate and pleasant than rude responses. Overall, the SDS was also perceived as more pleasant than the rude SDS. Both groups judged the likability of the SDS as moderate, with no differences between conditions. Also, conversational skill was not judged differently between conditions. The polite SDS was perceived as more accurate, showing more goodwill and integrity but not as having more expertise than the rude SDS. Rude responses were perceived as more face-threatening than polite responses. In the rude condition, most changes were aimed at strategies to mitigate FTAs (29 % on average), while, in the politeness condition, participants referred to providing more precise information and on changing or enhancing the employed voice.

Using a very clear and straightforward manipulation, we aimed to identify how the human principle of politeness is taken in consideration in SDS communication. In our setting, SDS answered to university freshmen's typical questions on student life. The only differences between our two experimental conditions were in the "how" part of the contribution, with rather rude or polite answers. The results mostly mirror those obtained with judgments of humans [12, 30]. These studies suggest that although both content and social information are transmitted with the same words, recipients seem to distinguish these two aspects; we routinely tease apart the "how" and "what" part of a contribution. In both previous studies and our present study, almost all social judgments, such as likability or the goodwill aspect of trustworthiness, were evaluated more positively in the polite condition. However, neither found evidence for differences in judgments of expertise, as a more content-related aspect of trustworthiness. However, in our study the accuracy of the system, which is also an arguably content-related aspect, was judged as higher for the polite system. Future research should ascertain whether this is a specific aspect of communication with an SDS or maybe a consequence of the specific realization of our conditions.

The results also indicate that advances on the technological side are expectations on the users' side. Politeness and adaptive communication require competent systems. Our between-subject design, which did not provide a direct comparison of rude and polite behavior, shows clearly that an SDS is judged relatively strictly according to its communication behavior. The users do not seem to be willing to be lenient simply because the system is not a human.

Our study has some limitations: one is that we had relatively young users assess the SDS. While those do represent typical users and the whole setting was ecologically valid for them, we cannot transfer the SDS evaluation results to other groups with less experience in technology (see the

Table 3: Change suggestions for each of the five categories from participants for each of the six SDS answers according to conditions (polite versus rude), reported in percentage.

	Answer 1		*Answer 2*		*Answer 3*		*Answer 4*		*Answer 5*		*Answer 6*		*Total*	
category	*polite*	*rude*	*polite*	*rude*	*polite*	*rude*	*polite*	*rude*	*polite*	*rude*	*polite*	*rude*	*polite*	*rude*
1 *(mitigate FTA)*	10	21	14	59	7	10	3	28	7	41	10	17	9	29
2 *(direct)*	3	0	3	0	3	0	0	0	3	0	10	10	4	2
3 *(voice)*	28	31	7	3	17	0	10	3	7	0	14	0	14	6
4 *(precise)*	7	3	21	0	21	21	14	3	21	7	3	0	15	6
5 *(no change)*	52	45	55	38	52	69	72	66	62	52	62	72	59	57

digital natives debate; [57]). Second, we simulated an indirect communication setting. Our participants did not interact with the SDS themselves; instead, they listened to responses of the SDS. There is evidence that a direct engagement with the SDS produces different results and absorbs some of the analytic abilities participants have shown in their evaluation of our ACURI [58].

One practical implication of this study might be drawn from our empirical evidence: regarding the first impression of an SDS and the evaluation of its acceptance and trustworthiness, the wording used by the SDS seems to have a considerable impact. Hence it might be worth engaging in creating more flexible and human-like technology. Communication principles such as politeness and alignment provide straightforward assumptions that might be embedded in technology and tested in experimental settings. Although the more attempts that are aimed at making assistants like ACURI become more human-like in their communication style, it is possible that such systems might also enhance miscommunication and misunderstanding. As an example, an SDS that is programmed to offer polite and indirect communication might produce responses that leave more room for interpretation and are thus less clear. In order to competently implement social communication factors in SDS, designers need to be aware how these principles come into play in different communication contexts. In this way, research into the social factors of human communication is highly relevant for the field of human-computer interaction.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by a grant awarded to the second author within the German Research Foundation's (DFG) Research Training Group GRK 1712: *Trust and Communication in a Digitized World*.

References

[1] N. Mavridis, "A review of verbal and non-verbal human-robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, no. 1, pp. 22–35, 2015.

[2] R. López-Cózar, Z. Callejas, D. Griol, and J. F. Quesada, "Review of spoken dialogue systems," *Loquens*, vol. 1, no. 2, Article ID e012, 2014.

[3] F. Manjoo, "The Echo from Amazon brims with groundbreaking promise," http://www.nytimes.com/2016/03/10/technology/the-echo-from-amazon-brims-with-groundbreaking-promise.html, 2016.

[4] M. F. McTear, *Spoken Dialogue Technology: toward The Conversational User Interface*, Springer Science & Business Media, 2004.

[5] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[6] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, *In the Shades of the Uncanny Valley: An Experimental Study of Human–Chatbot Interaction*, Future Generation Computer Systems, 2018.

[7] N. Dethlefs, H. Hastie, H. Cuayáhuitl, Y. Yu, V. Rieser, and O. Lemon, "Information density and overlap in spoken dialogue," *Computer Speech Language*, vol. 37, pp. 82–97, 2016.

[8] O. Vinyals and Q. Le, "A neural conversational model," arXiv preprint, arXiv:1506.05869. ISO 690, 2015.

[9] G. A. Linnemann and R. Jucks, "As in the question, so in the answer? Language style of human and machine speakers affects interlocutors' convergence on wordings," *Journal of Language and Social Psychology*, vol. 35, no. 6, pp. 686–697, 2016.

[10] C. I. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, MIT Press, Cambridge, UK, 2005.

[11] J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson, "Towards human-like dialogue systems," *Speech Communication*, vol. 50, no. 8, pp. 630–645, 2008.

[12] R. Jucks, G. A. Linnemann, F. M. Thon, and M. Zimmermann, "Trust the words: insights into the role of language in trust building in a digitalized world," in *Trust and Communication in a Digitized World*, B. Blöbaum, Ed., pp. 225–237, Springer International Publishing, 2016.

[13] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter, "Intelligent tutoring systems with conversational dialogue," *AI Magazine*, vol. 22, no. 4, pp. 39–51, 2001.

[14] R. E. Mayer, W. L. Johnson, E. Shaw, and S. Sandhu, "Constructing computer-based tutors that are socially sensitive: politeness in educational software," *International Journal of Human-Computer Studies*, vol. 64, no. 1, pp. 36–42, 2006.

[15] K. Porayska-Pomsta, M. Mavrikis, and H. Pain, "Diagnosing and acting on student affect: the tutors perspective," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1, pp. 125–173, 2008.

[16] M. Tai, I. Arroyo, and B. P. Woolf, "Teammate relationships improve help-seeking behavior in an intelligent tutoring system," in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 239–248, Springer, Berlin, Heidelberg, 2013.

[17] B. Whitworth, "Politeness as a social software requirement," *International Journal of Virtual Communities and Social Networking*, vol. 1, no. 2, pp. 65–84, 2009.

[18] B. M. McLaren, K. E. DeLeeuw, and R. E. Mayer, "Polite web-based intelligent tutors: can they improve learning in classrooms?" *Computers & Education*, vol. 56, no. 3, pp. 574–584, 2011.

[19] P. Brown and S. C. Levinson, *Politeness, Some universals in language Usage*, Cambridge University Press, Cambridge, UK, 1987.

[20] B. Brummernhenrich and R. Jucks, "Managing face threats and instructions in online tutoring," *Journal of Educational Psychology*, vol. 105, no. 2, pp. 341–350, 2013.

[21] T. Holtgraves, S. Ross, C. Weywadt, and T. Han, "Perceiving artificial social agents," *Computers in Human Behavior*, vol. 23, no. 5, pp. 2163–2174, 2007.

[22] A. De Angeli, W. Gerbino, E. Nodari, and D. Petrelli, "From tools to friends: where is the borderline?" in *Proceedings of the UM99 Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, pp. 1–10, Springer, Berlin, Germany, 1999.

[23] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction," *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, pp. 171–181, 2001.

[24] R. C. Mayer, J. H. Davis, and F. D. Shoorman, "An intergration model of organizational trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.

[25] S. Tseng and B. J. Fogg, "Credibility and computing technology," *Communications of the ACM*, vol. 42, no. 5, pp. 39–44, 1999.

[26] L. Hoffmann, N. C. Krämer, A. Lam-chi, and S. Kopp, "Media equation revisited: do users show polite reactions towards an embodied agent?" in *Intelligent Virtual Agents*, Z. Ruttkay, M. Kipp, A. Nijholt, and H. H. Vilhjálmsson, Eds., pp. 159–165, Springer, Berlin, Germany, 2009.

[27] A. De Angeli and S. Brahnam, "I hate you! Disinhibition with virtual partners," *Interacting with Computers*, vol. 20, no. 3, pp. 302–310, 2008.

[28] J. Culpeper, "Towards an anatomy of impoliteness," *Journal of Pragmatics*, vol. 25, no. 3, pp. 349–367, 1996.

[29] C. Nass, "Etiquette equality: exhibitions and expectations of computer politeness," *Communications of the ACM*, vol. 47, no. 4, pp. 35–37, 2004.

[30] B. Brummernhenrich and R. Jucks, "'He shouldn't have put it that way!' How face threats and mitigation strategies affect person perception in online tutoring," *Communication Education*, 2015.

[31] R. Jucks, L. Päuler, and B. Brummernhenrich, ""I need to be explicit: you're wrong": impact of face threats on social evaluations in online instructional communication," *Interacting with Computers*, vol. 28, no. 1, pp. 73–84, 2016.

[32] S. L. Jessmer and D. Anderson, "The effect of politeness and grammar on user perceptions of electronic mail," *North American Journal of Psychology*, vol. 3, no. 2, pp. 331–346, 2001.

[33] L. Dybkjær and N. O. Bernsen, "Usability issues in spoken dialogue systems," *Natural Language Engineering*, vol. 6, pp. 243–271, 2000.

[34] D. H. McKnight, "Trust in information technology," in *The Blackwell Encyclopedia of Management*, B. G. Davis, Ed., vol. 7 of *Management Information Systems*, pp. 329–331, Blackwell, Malden, MA, USA, 2005.

[35] R. C. Mayer and J. H. Davis, "The effect of the performance appraisal system on trust for management: a field quasi-experiment," *Journal of Applied Psychology*, vol. 84, no. 1, pp. 123-10, 1999.

[36] D. H. McKnight and N. L. Chervany, "Trust and distrust definitions: one bite at a time," in *Trust in Cyber-Societies*, R. Falcone, M. Singh, and Y.-H. Tan, Eds., pp. 27–54, Springer, Berlin, Heidelberg, 2001.

[37] M. D. Pickard, J. K. Burgoon, and D. C. Derrick, "Toward an objective linguistic-based measure of perceived embodied conversational agent power and likeability," *International Journal of Human-Computer Interaction*, vol. 30, no. 6, pp. 495–516, 2014.

[38] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean, "Linguistic alignment between people and computers," *Journal of Pragmatics*, vol. 42, no. 9, pp. 2355–2368, 2010.

[39] C. Torrey, A. Powers, M. Marge, S. R. Fussell, and S. Kiesler, "Effects of adaptive robot dialogue on information exchange and social relations," in *Proceedings of the HRI 2006: 2006 ACM Conference on Human-Robot Interaction*, pp. 126–133, USA, March 2006.

[40] J. J. Bradac, A. Mulac, and A. House, "Lexical diversity and magnitude of convergent versus divergent style shifting-: perceptual and evaluative consequences," *Language Communication*, vol. 8, no. 3, pp. 213–228, 1988.

[41] S. Gupta, M. Walker, and D. Romano, "How rude are you?: evaluating politeness and affect in interaction," *Affective Computing and Intelligent Interaction*, pp. 203–217, 2007.

[42] M. A. Walker, J. E. Cahn, and S. J. Whittaker, "Improvising linguistic style: social and affective bases for agent personality," in *Proceedings of the First International Conference on Autonomous Agents*, pp. 96–105, 1997.

[43] A. C. Graesser, "Learning, thinking, and emoting with discourse technologies," *American Psychologist*, vol. 66, no. 8, pp. 746–757, 2011.

[44] M. De Jong, M. Theune, and D. Hofs, "Politeness and alignment in dialogues with a virtual guide," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2008*, pp. 206–213, Portugal, May 2008.

[45] M. Thiebach, E. Mayweg-Paus, and R. Jucks, ""Probably true" says the expert: how two types of lexical hedges influence students' evaluation of scientificness," *European Journal of Psychology of Education*, vol. 30, no. 3, pp. 369–384, 2015.

[46] D. Bousfield and J. Culpeper, "Impoliteness, eclecticism and diaspora," *Journal of Politeness Research*, vol. 4, no. 2, pp. 161–168, 2008.

[47] N. Vergis and M. Terkourafi, "The role of the speakers emotional state in im/politeness assessments," *Journal of Language and Social Psychology*, vol. 34, no. 3, pp. 316–342, 2015.

[48] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, vol. 22, no. 11, pp. 1359–1366, 2011.

[49] W. R. Cupach and C. L. Carson, "Characteristics and consequences of interpersonal complaints associated with perceived face threat," *Journal of Social and Personal Relationships*, vol. 19, no. 4, pp. 443–462, 2002.

[50] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," *Natural Language Engineering*, vol. 6, no. 3, pp. 287–303, 2000.

[51] L. B. Larsen, "Assessment of spoken dialogue system usability-what are we really measuring?" in *Proceedings of the Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.

[52] S. Möller, P. Smeele, H. Boland, and J. Krebber, "Evaluating spoken dialogue systems according to de-facto standards: a case study," *Computer Speech & Language*, vol. 21, no. 1, pp. 26–53, 2007.

[53] F. Hendriks, D. Kienhues, and R. Bromme, "Measuring laypeoples trust in experts in a digital age the Muenster Epistemic Trustworthiness Inventory (METI)," *PLoS ONE*, vol. 10, no. 10, Article ID e0139309, 2015.

[54] J. C. McCroskey and J. J. Teven, "Goodwill, a reexamination of the construct and its measurement," *Communication Monographs*, vol. 66, no. 1, p. 90, 1999.

[55] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, 2015.

[56] A. Agresti, "A survey of exact inference for contingency tables," *Statistical Science*, vol. 7, no. 1, pp. 131–153, 1992.

[57] F. Salajan, D. Schonwetter, and B. Cleghorn, "Student and faculty inter-generational digital divide: fact or fiction?" *Computers and Education*, vol. 53, no. 3, pp. 1393–1403, 2010.

[58] G. A. Linnemann and R. Jucks, "Can I trust the spoken dialogue system because it uses the same words as i do?—influence of lexically aligned spoken dialogue systems on trustworthiness and unter satisfaction," *Interacting with Computers*, pp. 173–186, 2018.