Asia Pacific Journal of Corpus Research
Vol. 1, No. 2, 1-14

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/APJCR.2020.1.2.1

Gender-Based Differences in Expository Language Use: A Corpus Study of Japanese

Kevin Heffernan
(Kwansei Gakuin University)
Keiko Nishino
(Kwansei Gakuin University)

Heffernan, K., & Nishino, K. (2020). Gender-based differences in expository language use: A corpus s tudy of Japanese. *Asia Pacific Journal of Corpus Research*, *1*(2), 1-14.

Previous work has shown that men both explain and value the act of explaining more than women, as explaining conveys expertise. However, previous studies are limited to English. We conducted an exploratory study to see if similar patterns are seen amongst Japanese speakers. We examined three registers of Japanese: conversational interviews, simulated speeches, and academic presentations. For each text, we calculated two measures: lexical density and the percentage of the text written in kanji. Both are indicators of expository language. Men produced significantly higher scores for the interviews and speeches. However, the results for the presentations depend on age and academic field. In fields in which women are the minority, women produce higher scores. In the field in which men are the minority, younger men produced higher scores but older men produced lower scores than women of the same age. Our results show that in academic contexts, the explainers are not necessarily men but rather the gender minority. We argue that such speakers are under social pressure to present themselves as experts. These results show that the generalization that men tend to explain more than women does not always hold true, and we urge more academic work on expository language.

Keywords: Gender, Academic, Expository, Hypercorrection, Information Density

1. Introduction

The Merriam-Webster English dictionary defines mansplain as “(of a man) to explain something to a woman in a condescending way that assumes she has no knowledge about the topic.”[1] This word originated from Rebecca Solnit’s 2008 essay titled “Men explain things to me” (Solnit, 2014), which tells of her experiencing the bungling attempt of “Mr. Very Important” to lecture her on the merits of a newly published novel, only to find out that Solnit just so happens to be the author of the book that he is talking about. The stereotype behind the word mansplain can be decomposed into two parts: that men feel an urge to explain (even when they are not knowledgeable about the subject), and that they do so with an air of superiority. Our study is motivated by the first part. Our goal is to conduct an exploratory corpus-based study of gender-based differences in the use of expository language by Japanese speakers.

Specifically, we conduct three studies using different registers: conversational interviews, simulated public speeches, and academic presentations. The last genre is of particular interest to us as the objective of an academic presentation is to convey information in a limited amount of time. Gender-based differences in the amount of expository language used by academics in public forums

[1] Mansplain, Paragraph 1 of Merriam-Webster’s online dictionary. Retrieved from www.merriam-webster.com/dictionary/mansplain.

has to our knowledge not been explored. Given the limited flexibility of the discourse, we naively hypothesize that academic presentations are void of gendered-based differences in expository language. In contrast, based on the literature reviewed we expect to find gendered-based differences in the conversational interviews.

Our approach differs from most modern work on gender and language in that we take the binary gender categories of female and male as the starting point for our comparative analysis. We take a quantitative approach informed by the structural methods of analysis of corpus linguistics. Such an approach, when applied to the study of language and gender, has its own merits (Hultgren, 2008). A quantitative analysis is particularly ideal for an exploratory approach to finding potentially meaningful gender-based differences. Such differences are often obvious, but they can also be subtle in nature, such as the phenomenon that we examine in this paper. Corpus-based analysis makes for ideal testing grounds since large volumes of speech produced by many speakers increases the sensitivity of our tools to such subtle differences that would otherwise remain undetected and unreported. However, such differences, as we argue in our concluding remarks, stem from broad, sweeping social ideologies of gender in academic contexts. Thus, once such differences have been discovered, future work may then build on these preliminary steps through more critical, fine-grained analysis.

In the next section we review the limited work on gender-based differences in expository language. We then present our corpus study of gender-based differences in Japanese. We conclude by drawing connections between our results and sociolinguistic work on mobility and social class.

2. Background

2.1. Expository Language

Expository language refers to discourse designed to convey information, such as a detailed explanation. Expository language contrasts with the narrative language of fiction writing, which is used to tell a story. One clear indicator of expository language is noun usage rate (Biber, 1988; Biber, Conrad, Reppen, Byrd, & Helt, 2012; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Ravid & Berman, 2010). For example, news reports and academic writing contain on average almost four nouns for every verb; in contrast, conversational speech contains roughly one noun for every verb (Biber et al., 1999, p. 65). Biber (1988) explains this difference as reflecting the degree of “informational density” (p. 107). Information density refers to the rate of conveying meaning. Integrating meaning-bearing words such as nouns into the text through devices such as compounding (1) and adjunct phrases (2) leads to compressed expressions that convey meaning in few words (Biber & Gray, 2010). The expressions (1) and (2) are more economical than their equivalents given in (3) and (4), and therefore they have higher information density.

1) alcohol consumption
2) a surrogate for the country as a whole
3) the process of consuming alcohol
4) a surrogate that represents the country as a whole

Noun usage rate is not the only source of difference between expository language and more narrative language. Biber (1988) conducted a multivariate analysis on a general corpus of spoken and written English, and isolated five main dimensions. His first dimension, which he titled involved versus informative production, showed that increased information production correlated with increased noun usage, word length, and type/token ratio. Nouns convey referential meaning, and

therefore an increased rate of noun usage corresponds to increased informational density. Similarly, an increased type/token ratio reflects increased variety in lexical choice, corresponding to a more precise presentation of information.

Biber et al. (2012) used the same multidimensional analysis to examine the characteristics of academic speaking and writing. They found that written discourse (textbooks, course packs, etc.) was highly differentiated from spoken discourse (classroom teaching, office hours, etc.) by the frequent use of nouns, longer words, and higher type/token ratios. Biber et al. (2012) interpret these results as "indicating that the written registers are extremely informational in purpose" (p. 143).

2.2. Gender-based Differences in Expository Language

Previous work on gender-based differences in the use of expository language is limited. However, one relevant study is Sauntson (2012). Sauntson studied gender differences that emerged during group work interactions of 11-year-old to 14-year-old students. She conducted a variety of analyses, including a discourse analysis focused on the functions performed by utterances in relation to each other. She observed notable differences between the girls and the boys with regards to the rate that certain types of exchanges were used. An exchange is a unit of discourse composed of several moves, such as opening, answering, and acknowledging. Sauntson classified each exchange as one of nine types, but more than 80% of the exchanges were accounted for by just three types: elicit information from the interlocutor, inform the interlocutor of something, and direct the interlocutor to do something. When she broke down the data by gender dyad, she found that 42.7% of the boy-boy dyad exchanges were the inform type, notably higher than the proportion of the girl-girl dyads (31.6%), or the mixed dyads (37.1%). Further decomposition of the exchanges into their component moves revealed that the boys use informative moves at a rate of 82.7%, compared to only 58.0% by girls. Sauntson interprets these differences as showing that "boys are confident about asserting their own individual knowledge within the group and about valuing it as a learned resource which can be used by others" (Sauntson, 2012, p. 77). She also observed that the boys seldom endorsed each other's informative moves with a response. Rather, an informing move was often followed with a competing informing move as the boys "continually engaged in struggles over who holds the most power and status within the group" (Sauntson, 2012, p. 77). Finally, Saunston also observed that the boys make frequent use of the first-person pronouns during these exchanges, suggesting that the boys are more concerned with their individual achievements than with the group's collaborative achievements.

Coates (2016) similarly observed that during male-male dyads, speakers would hold the floor for a considerable length of time as they explained something. She interprets this act as "playing the expert" (Coates, 2016, p. 134), and notes that it is a game played most commonly by male speakers. Thus, we see limited evidence of men using more expository language than women. Furthermore, a desire to establish themselves as capable seems to underlie their desire to present knowledge. The observations of Sauntson and Coates are based on the fluid back and forth of discourse, during which interlocutors capitalize on their subject agency to continually negotiate their gender identity. In this way, these studies and many others demonstrate the emergence of social gender through discourse.

As for studies on gender-based differences in Japanese, we were unable to find a precedent. However, Matsuda (2014) reports a greater noun to verb ratio for male speakers compared to female speakers in the Ozaki questionnaire study on honorific language use in Japanese. Matsuda explains the gender difference as a consequence of the greater use of honorific auxiliary verbs by female speakers. However, his results are also consistent with the observation that men in general use more expository language than women in conversation.

In this study, we examine gender differences in the use of expository language in three registers: sociolinguistic interviews, simulated short public speeches, and academic conference presentations. We present the details of our data in the methodology section below.

3. Methodology

In this section we first introduce the three data sources that we use in this study, and then we explain how we measure information density.

3.1. The Three Data Sources

We drew our data from the *Corpus of Kansai Vernacular Japanese* and the *Corpus of Spontaneous Japanese*.[2, 3] Both of these corpora are parsed and tagged with part of speech information using a process of manual training and automatic parsing with the Mecab parser.[4] We used data from three registers, the conversational interviews from the *Corpus of Kansai Vernacular Japanese*, and the simulated public speeches and the academic presentations from the *Corpus of Spontaneous Japanese*. Each dataset is described in turn.

The *Corpus of Kansai Vernacular Japanese* consists of 160 conversational interviews (hereafter, conversations) conducted by university students attending a private university in the Kansai area of Japan with either a family member or a close acquaintance of the student in a casual setting. The conversations last on average 55.19 minutes (*SD* = 5.13). Both the students and the interviewees self-reported that they were native speakers of the Kansai dialect. The interviewees were selected so that they were balanced for age and gender. Approximately half of the interviewees (*N* = 82) were female. This study only examines the speech produced by the interviewees (a total of 937,000 words). The utterances of the students were systematically excluded.

The simulated public speeches (hereafter, speeches) were recorded in a phonetics studio at the National Institute for Japanese Language and Linguistics, and produced by 1,714 paid participants, of which approximately half (*N* = 908) were female. Each participant spoke for an average of 11.62 minutes (*SD* = 2.03 minutes) on everyday topics, such as their happiest memory, producing a total of 3.60 million words. The speakers spoke in front of a very small audience of between one person and ten people.

The academic presentations (hereafter, presentations), were produced by 963 presenters who were recorded live during conferences covering topics in the fields of engineering, social science, and the humanities. Approximately one fifth (*N* = 164) of the speakers were female. The average duration of the recordings is 16.68 minutes (*SD* = 6.89). The recordings include 19 presentations that were not the standard academic presentation format. These texts were removed, leaving 944 texts consisting of a total of 2.85 million words.

3.2. Calculating Information Density

We used two measures to calculate the information density of the 2,818 texts: the lexical density of the text and the proportion of transcription that was written in kanji. As not only nouns but also verbs and adjectives convey information, we opted to use lexical density instead of a measure based only on nouns. We calculated the lexical density of a text as the proportion of words that are nouns, verbs, and adjectives. We excluded from the count clitic-like nouns such as the *-sa* in *naga-sa* 'length,' which consists of *naga-* 'long' and the nominalizer *-sa*. We counted grammatical nouns such as pronouns as not lexical. The mean lexical density for the conversations is 32.52% (*SD* = 2.23%), for the speeches is 40.56% (*SD* = 4.13%), and for the presentations is 47.82% (*SD* = 4.37%).

The second measure that we used is the proportion of kanji used in the transcription. Increased kanji usage reflects increased Sino-Japanese vocabulary usage. Sino-Japanese vocabulary is to the

[2] This corpus is available for download from https://sites.google.com/view/kvjcorpus.

[3] This corpus may be accessed online at https://pj.ninjal.ac.jp/corpus_center/csj/.

[4] This parser may be obtained from http://taku910.github.io/mecab/.

Japanese language as Latinate vocabulary is to the English language: Sino-Japanese constitutes the (older) technical and learned vocabulary of the language (Itō & Mester, 1999). However, the distinction in Japanese is more conscious since it is reflected in the writing. Specifically, Sino-Japanese words are written only in kanji, whereas native words are written in a combination of kanji and hiragana script. Each character, excluding punctuation, was classified as either kanji or not, and the proportion of the transcription that was kanji characters was calculated as a percentage. The average proportion of the transcript written in kanji for conversations is 21.52% (*SD* = 2.41%), for the speeches is 27.16% (*SD* = 4.17%) and for the academic presentations is 32.63% (*SD* = 5.20%).

As the unit of measurement for both of our measures is percentage, we can combine the two measures without any scaling. We calculate the average of the two measures for each text and call this score the information density score of that text. The average information density score for the conversations is 27.02% (*SD* = 2.11%), for the speeches is 33.86% (*SD* = 3.92%), and for the presentations is 40.23% (*SD* = 4.57%). This result shows that the three registers are differentiated by information density as expected, with a higher information density for the more expository texts.

4. Results

4.1. Gender Differences: Initial Results

In this section, we test for gender-based differences in information density in our three datasets. We do so with independent-samples Student's t-tests that compare the average information density score of the two genders.

Our test on the conversations dataset showed a significant difference in information density between the women (M = 26.70, SD = 1.99) and men (M = 27.36, SD = 2.20); $t(158) = -1.97$, $p = 0.050$. Men, on average, produced texts with a higher information density than the women. This result is consistent with the observations of Sauntson (2012) and Coates (2016), who both reported male speakers using more expository speech than women.

Our test on the speeches dataset also showed a significant difference in information density between the women (M = 33.59, SD = 3.73) and men (M = 34.11, SD = 4.14); $t(1{,}172) = -2.41$, $p = 0.016$. Again, men produced texts with a higher information density than the women.

Examining the presentations data shows a different pattern. Once again, our test showed a significant difference in information density between the women (M = 41.5, SD = 4.53) and men (M = 40.1, SD = 4.50); $t(946) = 3.66$, $p < 0.001$. However, in this dataset the men produced datasets with a lower information density than the women.

The presentations result is surprising. The conversations and speeches datasets show the expected pattern based on previous research: men on average produce texts with higher information density than women. However, for all three datasets, the group difference is notably smaller than the standard deviations, showing that the intragroup variation is larger than the intergroup variation. The gender-based differences may be a consequence of unaccounted-for factors, such as the gender of the interviewer, the topic of the speech, or the academic field of the presentation. Fortunately, the metadata for the corpora lists a wide range of information for each of the speakers, including interviewer gender, the speech topic, and academic field. In the following sections, we use this information to refine our analysis and shed light on the gender-based differences in information density.

4.2. Reanalysis of the Conversations Data

The following meta information was used in the reanalysis of the conversations dataset: speaker

gender (1: female, *N* = 82; 0: male, *N* = 78)[5]; speaker age (1: 15 to 23 years old, *N* = 46; 2: 24 to 39 years old, *N* = 36; 3: 40 to 59 years old, *N* = 39; 4: 60 to 86 years old, *N* = 39); and whether or not the speaker has post-secondary education (1: yes, *N* = 88; 0: no, *N* = 72). Following previous work on gender-based differences in information density (Coates, 2016; Sauntson, 2012), the gender of the research assistant was recorded based on the makeup of the interview dyad, either female-female (*N* = 61), male-male (*N* = 44), or mixed (*N* = 55).

It is possible that the above-mentioned factors account for the gender-based differences in information density in the conversations dataset. In order to address this possibility, we first model the variation in information density without gender as a factor. We then add gender and interactions with gender, if any, to the base model. Finally, we compare the base model with the new model using hierarchical regression analysis. If the difference in R-square between the base model and the new model is statistically significant, then we conclude that there are significant gender differences in information density above and beyond that which has been accounted for by other factors.

Table 1 shows the results of the hierarchical regression analysis for the conversations. The regression analysis of the base model yielded only one statistically-significant factor, dyad. Male-male dyads show the highest information density (*M* = 27.58, *SD* = 1.92), followed by mixed dyads (*M* = 27.12, *SD* = 2.30), and then female-female dyads (*M* = 26.53, *SD* = 1.99). Adding gender improves R-square by only 0.1%, which is a non-significant difference; $F(2,154) = 0.19$, $p > 0.05$. As there were not any significant interactions with gender, the interaction terms were left out of the new model. These results are consistent with previous research on gender and language that has shown that gender-based linguistics differences are attenuated in mixed-gender dyads (Mulac, Wiemann, Widenmann, & Gibson, 1988).

However, and perhaps more importantly, the differences between the dyads are relatively small. Without further research, we can only speculate. With that caveat, we suggest that perhaps the lack of notable differences stems from the removal of the speakers from their usual communities of practice. Within the socially limited constraints of the conversational interview, speakers may be less motivated to perform being an expert.

Table 1. Hierarchical Regression Analysis Comparing the Base Model Against the Model with Gender, for the Conversations Dataset

Independent variables	Base model	Model with gender
Age (1: student; 4: older adult)	-0.05	-0.05
Dyad		
female-female	baseline	baseline
mixed	0.60	0.77
male-male	1.10**	1.36
Post-secondary education (1: yes; 0: no)	-0.43	-0.45
Gender (1: female; 0: male)		0.25
R^2	0.051	0.052
R^2 change		0.001

**$p < 0.01$.

4.3. Reanalysis of the Speeches Data

The following meta information was used in the reanalysis of the speeches data: speaker gender (1: female, *N* = 906; 0: male, *N* = 806); speaker age (1: under 30, *N* = 502; 2: 30 to 44, *N* = 539; 3: 45 to 59, *N* = 393; 4: 60 years or older, *N* = 280); audience size (1: one or two, *N* = 519; 2: three, *N* = 878; 3: four, *N* = 255; 4: five to ten, *N* = 60); the theme of the speech (history or news, *N* = 256; life experiences such as a

[5] The contents of the brackets indicates the value the category was recoded as, and the number of speakers in that category.

happy moment, N = 446; objective reporting such as an explanation of how to do something, N = 440; personal opinion, N = 351); the amount of preparation by the speaker (1: none, N = 33; 2: notes, N = 1358; 3: script, N = 307); and whether or not the speaker has completed post-secondary education (0: no, N = 574; yes: N = 1,134).

The authors of the corpus impressionistically evaluated the speeches on several five-step Likert scales. The responses were quantified as integers ranging from one to five. The scales that we use, their means, and standard deviations are as follows: spontaneity (M = 4.02, SD = 1.08), technical complexity (M = 1.97, SD = 1.13), speaking speed (M = 3.12, SD = 0.79), and pronunciation clarity (M = 3.12, SD = 0.98). Each speaker was also impressionistically evaluated for their speaking proficiency on a four-step scale, which we quantified as integers ranging from one to four (M = 2.01, SD = 0.74).

Table 2 shows the results of the hierarchical regression analysis for the speeches. The regression analysis of the base model contains multiple statistically-significant factors, and has an R-squared value of 36.6%. The variance inflation factors of the predictors were checked to determine if multicollinearity was a problem. As none of the factors were greater than 1.3, we included all of factors.

One of the factors, age, showed an interaction with gender. Adding gender and the gender-age interaction to the model increased the R-squared value by 1.5%, a small but statistically significant improvement; $F(2,1,433) = 26.28$, $p < 0.001$. The multiple regression analysis shows that information density increases as the following factors increase: age, education level, preparation, speaking proficiency, and amount of technical language. On the other hand, information density decreases as the following factors increase: clarity, speaking speed, and spontaneity. With regards to the topic of the speech, the objective topics (news, history, objective reporting, M = 35.1%, SD = 4.29%) showed greater information density than the subjective topics (personal opinion and life experiences, M = 32.8%, SD = 3.34%).

As mentioned, the male speakers produced texts with higher information density than the female speakers. This gender effect interacts with age such that the younger female speakers produce texts with higher information density than the younger male speakers, the opposite of the general result. However, this interaction depends on the theme of the talk. The interaction is clearest for objective-themed talks (Figure 1). When the speakers spoke about objective themes, the youngest female speakers produced texts with greater information density than those of the youngest male speakers. In contrast, when the speakers spoke about subjective themes, the younger female speakers produced texts that were of a comparable information density to those of the younger male speakers (Figure 2).

In summary, men tend to use more expository language than women. However, the exception to this generalization is young women talking about a non-personal topic. In such a case, young women tend to use more expository language than men of a comparable age.

The relationship between age and expository information is worth pointing out. Figure 1 and Figure 2 both show that as the speaker's age increases, the use of expository language increases. Such a correlation between language and language use indicates either an ongoing change in the language, or age grading (Bailey, 2002). Age grading refers to linguistic variation that correlates with a particular phase in life and repeats from one generation to the next. Given the lack of a correlation with age in the conversations data, an ongoing change in the Japanese language seems unlikely. In that case, we are left with the possibility that for some reason older speakers use more expository language than younger speakers when giving a public speech. We return to this point in the discussion.

In the next section, we reexamine the presentations data. Since the content of academic presentations is objective in nature, we expect results similar to Figure 1.

4.4. Reanalysis of the Presentations Data

The following meta information was used in the reanalysis of the presentations data: speaker gender (1: female, N = 164; 0: male, N = 780); speaker age (1: either under 25, N = 194; 2: 25 to 29, N =

218; 3: 30 to 39, *N* = 268; 4: 40 years or older, *N* = 253); audience size (1: less than 30, *N* = 184; 2: 30 to 49, *N* = 244; 3: 50 to 99, *N* = 251; 4: 100+, *N* = 258); academic field (engineering, e.g., artificial intelligence, natural language processing, *N* = 652; social science, e.g., linguistics, psychology, *N* = 194; Japanese language education, *N* = 88); and whether or not the speaker has completed post-graduate education (0: no, *N* = 283; 1: yes, *N* = 653).

Table 2. Hierarchical Regression Analysis Comparing the Base Model Against the Model with Gender, for the Speeches Dataset

Independent variables	Base model	Model with gender
Age (1: under 30; 4: over 60)	0.39***	0.81***
Audience (1: one or two; 4: five to ten)	-0.18	-0.14
Clarity (1: no at all; 5: very much)	-0.25**	-0.16†
Post-secondary education (1: yes; 0: no)	0.50**	0.27
Preparation (1: none; 3: script)	0.42†	0.37†
Proficiency (1: no at all; 5: very much)	0.38**	0.26*
Speed (1: very slow; 5: very fast)	-0.50***	-0.51***
Spontaneity (1: not at all; 5: very much)	-1.46***	-1.47***
Technical language (1: not at all; 5: very much)	0.51***	0.49***
Theme		
News, history	baseline	baseline
Objective reporting	-1.05***	-1.09***
Personal opinion	-1.90***	-1.96***
Life experiences	-2.14***	-2.22***
Gender (1: female; 0: male)		1.37**
Gender x Age		-0.83***
R^2	0.366	0.381
R^2 change		0.015***

†$p < 0.1$, *$p < 0.5$, **$p < 0.01$, ***$p < 0.001$.

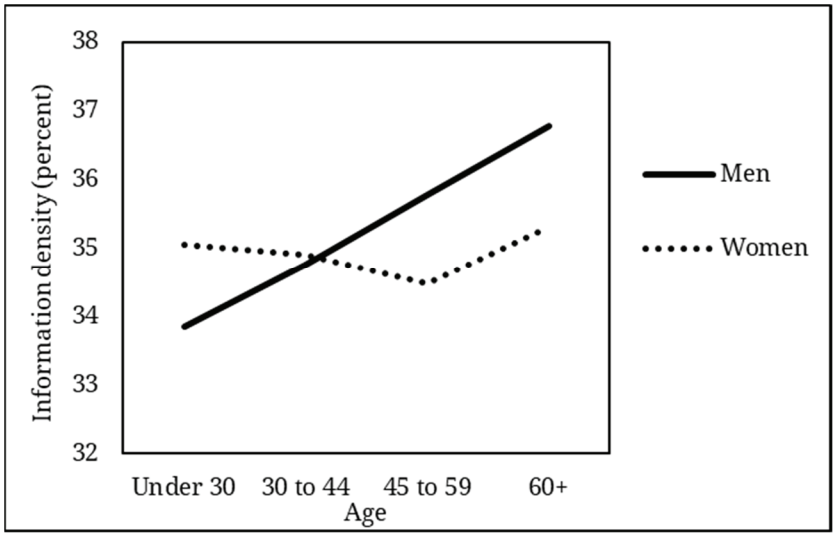

Figure 1. Information Density by Age and Gender, for the News, History, and Objective Reporting Speeches

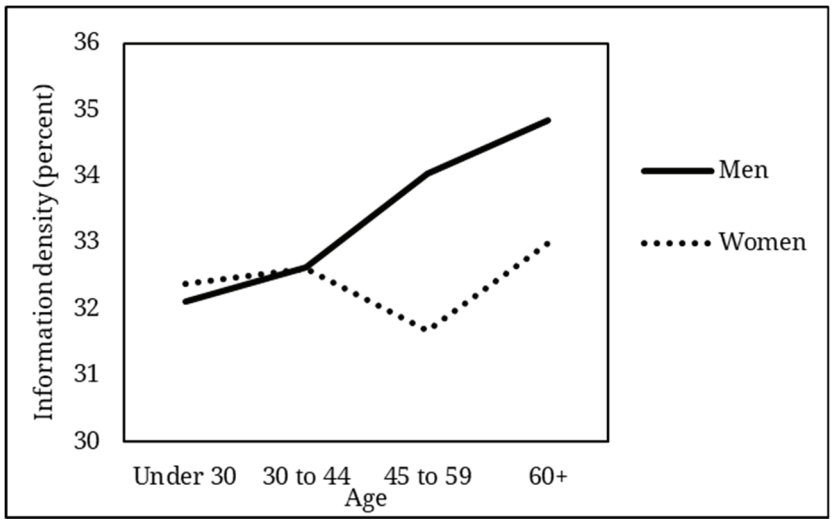

Figure 2. Information Density by Age and Gender, for the Personal Opinion and Life Experiences Speeches

Similar to the speeches data, each text in the presentations has been evaluated on several five-step Likert scales. The following scales were used in the analysis: spontaneity (M = 3.38, SD = 1.27), technical complexity (M = 3.30, SD = 0.97), speaking speed (M = 3.24, SD = 0.82), and pronunciation clarity (M = 3.47, SD = 1.01).[6]

Table 3 shows the results of the hierarchical regression analysis for the presentations. The base model has an R-squared value of 34.6%, and none of the predictors have a variance inflation factor over 1.75.[7] None of the predictors showed an interaction with gender, so only gender was added to the second model. Adding gender improved the R-squared value by 1.0%, a small but statistically significant value; $F(2, 758) = 12.12$, $p < 0.001$.

From Table 3, we see that age, field, and gender all strongly correlate with information density. A closer examination of the relationship between these three variables and information density reveals a complicated picture. In the case of the engineering and social science presentations, the female speakers on average produce texts with higher information density then their male counterparts of the same age (Figure 3). However, in the case of the Japanese language education speeches, we once again see a crossover pattern: The younger males on average produced texts with greater information density scores than their female counterparts (Figure 4).

Returning to relationship between age and expository language, we see notably different results than those of the speeches data. In general, as the age of the speaker increases, the use of expository language decreases—the opposite trend from the previous section. Such a trend rules out the possibility of an ongoing language change, leaving age grading as the probable explanation. However, why do the older speakers use more expository language than younger speakers when giving simulated speeches but use less expository language than younger speakers when giving academic presentations? We attempt to explain these seemingly contradicting results in the Discussion.

---

[6] The following factors are missing between eight and 25 items from the metadata: age, audience size, education, and academic field. The following factors are missing between 86 and 104 items: pronunciation clarity, speaking speed, and technical complexity. Of the presentation texts, 29 were labelled as workshop or research meeting. As we are unable to determine the field of the presentations, we recoded their academic field as a missing data item.

[7] We also initial considered three other measures: amount of preparation, speaking proficiency, and years of experience giving academic presentations. However, a variance inflation factor analysis indicated a high degree of multicollinearity. Specifically, amount of preparation inversely correlated with spontaneity, speaking proficiency correlated with spontaneity, and years of experience correlated with age. Running various multivariable hierarchical regression analyses showed that the model that we present has the greatest R^2 value.

Table 3. Hierarchical Regression Analysis Comparing the Base Model Against the Model with Gender, for the Presentations Dataset

Independent variables	Base model	Model with gender
Age (1: under 25; 4: over 40)	-1.26***	-1.27***
Audience (1: <30; 4: 100+)	0.39**	0.38**
Clarity (1: no at all; 5: very much)	-0.28*	-0.35**
Post-graduate education (1: yes; 0: no)	-0.72*	-0.78*
Speed (1: very slow; 5: very fast)	-0.01	-0.03
Spontaneity (1: not at all; 5: very much)	-1.23***	-1.15***
Technical language (1: not at all; 5: very much)	0.06	0.13
Field		
Engineering	baseline	baseline
Japanese lang. education	-2.60***	-2.62***
Social science	-2.97***	-3.16***
Gender (1: female; 0: male)		1.26***
R^2	.346	.356
R^2 change		.010***

†$p < 0.1$, *$p < 0.5$, **$p < 0.01$, ***$p < 0.001$.

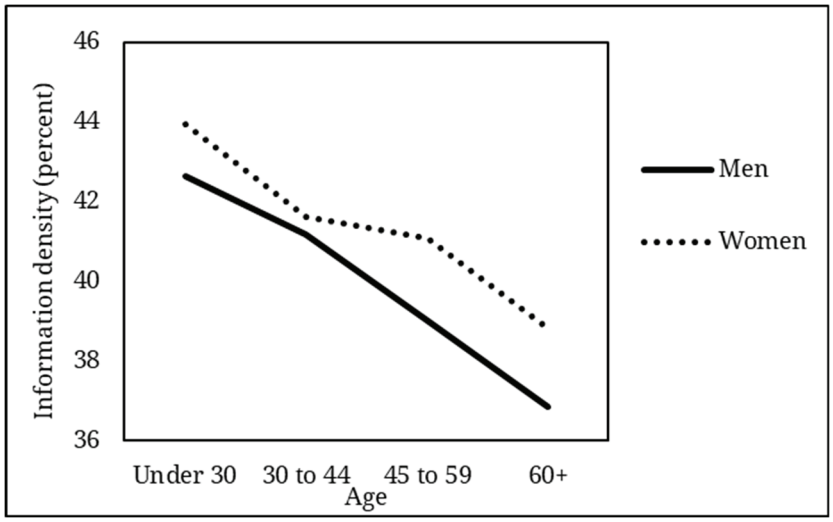

Figure 3. Information Density by Age and Gender, for the Engineering and Social Sciences Presentations

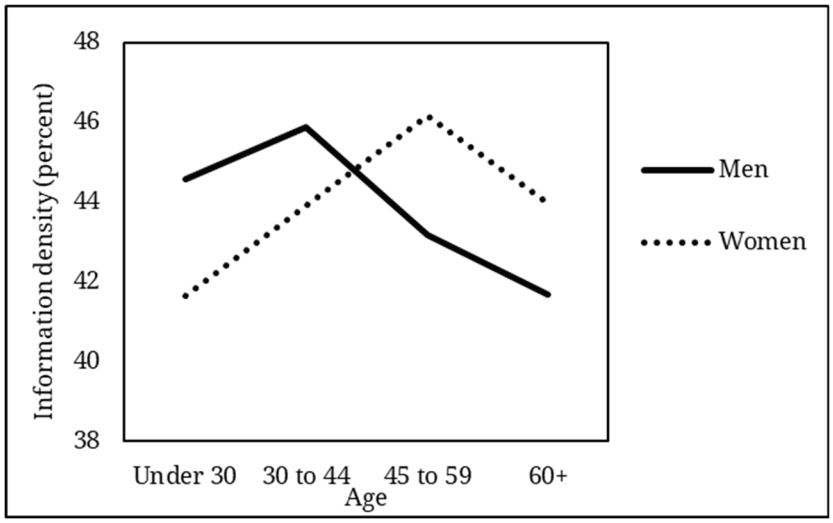

Figure 4. Information Density by Age and Gender, for the Language Education Presentations

5. Discussion and Conclusions

Generally speaking, men seem to want to explain things more than women. This stereotype has motived the English slang word "mansplain." Previous work examining this phenomenon has suggested that the motivation behind this desire to explain seems to be twofold (Coates, 2016, Sauntson, 2012). First, explanations are seen as a valuable resource to be used by others to further group achievement. Second, explanations are used to index power and status within the group by presenting oneself as an expert. Thus, explainers may be more concerned with individual achievements than with the group's collaborative achievements. Particularly relevant to our study is the suggestion that explainers desire to establish themselves as capable within the group.

In this study, we examined gender-based differences in the use of expository language in three registers of Japanese: conversational interviews, simulated public speaking, and academic presentations. For each text, we calculated an expository language score. This score was based on the lexical density of the text and the percentage of the text written in kanji.

In general, men produced significantly higher scores for the conversations and the speeches. However, young women talking about a non-personal topic showed the opposite trend compared to men of a comparable age.

The results for the presentations depend on age and academic field. In the fields of engineering and social science, women consistently produced significantly higher scores than men of a comparable age. In language education, younger men produced higher scores than younger women, whereas the opposite pattern was seen amongst older speakers.

Building on the results of previous research by Coates and Sauntson, we speculate that women in the sciences and young men in language education may desire to present themselves as experts to a greater degree than their gender counterparts. If so, then why? According to a 2017 Japanese government report[8] on gender and society, women in the sciences are the minority: They make up 41.6% of graduate students in social sciences, and only 12.3% in engineering. In contrast, women make up 49.5% of graduate students in education, a value which would certainly be higher if we restricted it to language education only, as the corpus data does. Thus, our results show that in academic contexts the explainers are not necessarily men but rather the gender minority.

The greater use of expository language by the gender minority resembles hypercorrect linguistic behavior. Work on sociolinguistic variation beginning with Labov (1966) has shown that speakers who moved up in social class from their childhood adjust their usage of certain linguistic features to sound more like the class that they joined. Interestingly, such speakers tend to hypercorrect—that is, overapply the features that they are adopting. Hypercorrection occurs as reduced usage of non-standard pronunciation variants associated with working class speech, such as pronouncing the voiced fricative /th/ in the word *this* as a voiced stop [d]. Such hypercorrection is most prominent among lower middle-class speakers who have moved up from the working class. Such speakers tend to underuse non-standard pronunciation variants and overuse standard pronunciation variants, compared to other lower middle-class speakers. Crucially, socially stable lower middle-class speakers do not show hypercorrection. Finally, Labov (1972, p. 141) notes that hypercorrectness is stronger among women than men.

Women who have entered the sciences may be behaving similar to social upwardly mobile individuals. Due to their minority status, they may feel social pressure to present themselves as experts. A newspaper article[9] titled "Scholar calls Japan's gender problem 'human disaster'" gives some insight into this social pressure with reference to the situation at the University of Tokyo, where women are

8 Page 92 of the Cabinet Office's 2017 *Danjo Kyoudou Sankaku Hakusho* [White Paper on Gender Equality]. Retrieved from http://www.gender.go.jp/about_danjo/whitepaper/h29/

9 *The Japan Times*, July 3rd, 2019, page 3.

a clear minority:

> The 70-year-old scholar... pointed out that less than 20 percent of students entering the University of Tokyo... are female. This is despite the evidence that female applicants have higher standard scores than their male competitors.
>
> [The scholar] said one reason for the discrepancy is the gender discrimination that is inherent in the education investment decisions made by Japanese parents. Traditionally, sons are expected to attend university while daughters will often be enrolled in a two-year junior college.
>
> Even female students at [the University of Tokyo] feel compelled to conform to Japanese social norms. [The scholar] spoke of how female students hide the fact that they were able to enter the prestigious institution in order to not scare off male students from other universities [since males dislike dating females who have accomplished greater academic achievements than themselves].

Japanese society seems to expect that women are not capable of the intellectual endeavor required of a University of Tokyo student. Given the low enrollment in the sciences in general, such gender bias is most likely ubiquitous. Against this social backdrop, it is no wonder that women in the sciences feel obliged to present themselves as an expert to a greater extent than men.

Without further data, explaining the pattern produced by the language education presenters is difficult. It may be that young men in language education are a clear minority, and therefore feel pressure to present themselves as an expert. On the other hand, it may be that older women feel pressure to present themselves as experts, and therefore use more expository language than their male counterparts.

We are also unable to explain why young women would feel pressure to use more expository language than men when giving a simulated public speech about an objective subject such as history. These women may also feel pressure to appear like an expert given the social prejudice against them.

The opposing age trends seen in the simulated speeches and the academic presentations also needs to be considered in light of our performance of expertise hypothesis. First consider the simulated speeches. Older speakers working outside of academia are the least familiar with the university environment. Given this lack of familiarity, they could feel greater pressure to perform than younger speakers. Such an explanation fits the academic presentations results. The older an academic becomes, the more experience her or she will have giving academic presentations, and therefore the more comfortable he or she will be when doing so.

In conclusion, our study has shown that the generalization that men tend to explain more than women does not always hold true. We have presented at least scenario in which the opposite is true: academic presentations in the sciences. We have argued that such women are under social pressure to present themselves as experts. The next logical step is ethnographic research designed to shed light on the possible underlying motivations for greater use of expository language. Such research would ideally focus on individual speakers and examine questions such as to what extent are speakers conscious of their relative use of expository language and to what extent do speakers feel pressure to present themselves as an expert.

Acknowledgments

This research was sponsored by Grant in Aid for Scientific Research (C) 17K02761 from the Japan Society for the Promotion of Science. We thank the participants of the NWAV-AP 2020 Tokyo conference for feedback.

References

Bailey, G. (2002). Real and apparent time. In Chambers, J. K., Trudgill, P., & Schilling-Estes, N. (Eds.), *The Handbook of Language Variation and Change* (pp. 312-332). Oxford: Blackwell.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2012). Speaking and writing in the university: A multidimensional comparison. In Biber, D., & Reppen, R. (Eds.), *Corpus Linguistics: Vol. 3. Varieties* (pp. 129-164). London: Sage.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), 2-20.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Essex: Longman.

Coates, J. (2016). *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. London: Routledge.

Hultgren, A. K. (2008). Reconstructing the sex dichotomy in language and gender research: Some advantages of using correlational sociolinguistics. In Harrington, K., Litosseliti, L., Sauntson, H., & Sunderland, J. (Eds.), *Gender and Language Research Methodologies* (pp. 29-42). Basingstoke: Palgrave Macmillan.

Itō, J., & Mester, A. (1999). The phonological lexicon. In Tsujimura, N. (Ed.), *The Handbook of Japanese Linguistics* (pp. 62-100). Malden, MA: Blackwell.

Labov, W. (1966). The effect of social mobility on linguistic behavior. *Sociological Inquiry*, *36*, 186-203.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

Matsuda, K. (2014). Katitaiso kaiseki no daikibo gengo chousa deeta e no ouyou: Okazaki keigo chousa paneru deeta ni okeru meishi, daimeishi, doushi no soutai hindosuu ni taisuru washa seibetsu kouka no kenshou [Analyzing large-scale POS-tagged language survey data: A case of sex effects on the noun-to-verb ratio in the OSH panel data]. *NINJAL Research Papers*, *7*, 151-165.

Mulac, A., Wiemann, J. M., Widenmann, S. J., & Gibson, T. W. (1988). Male/female language differences and effects in same-sex and mixed-sex dyads: The gender-linked language effect. *Communication Monographs*, *55*(4), 315-335.

Ravid, D., & Berman, R. A. (2010). Developing noun phrase complexity at school age: A text-embedded cross-linguistic analysis. *First Language*, *30*(1), 3-26.

Sauntson, H. (2012). *Approaches to Gender and Spoken Classroom Discourse*. Basingstoke: Palgrave Macmillan.

Solnit, R. (2014). *Men Explain Things to Me*. Chicago, Illinois: Haymarket Books.

THE AUTHORS

Kevin Heffernan is a Professor of linguistics and culture at Kwansei Gakuin University. He earned his Ph.D. in linguistics in 2007 at the University of Toronto, Canada. His main area of interest is vernacular Japanese. He is the creator of the *Corpus of Kansai Vernacular Japanese*.

Keiko Nishino is an Associate Professor at School of Policy Studies, Kwansei Gakuin University. She earned her MA in international relations in 1981 from the School of Advanced International Studies, Johns Hopkins University, USA. Her research interests include gender and development, international cooperation through social business, and evaluation of development interventions.

THE AUTHORS' ADDRESSES

First and Corresponding Author
Kevin Heffernan
Professor
School of Policy Studies
Kwansei Gakuin University
2-1 Gakuen, Sanda, Hyogo 669-1337, JAPAN
E-mail: kevin@kwansei.ac.jp

Co-author
Keiko Nishino
Professor
School of Policy Studies
Kwansei Gakuin University
2-1 Gakuen, Sanda, Hyogo 669-1337, JAPAN
E-mail: cia49965@kwansei.ac.jp

Received: 28 October 2020
Received in Revised Form: 7 December 2020
Accepted: 16 December 2020