A Comparative Corpus Study on the Use of Personal Pronouns in Japanese High School English Textbooks

Akiko Hagiwara[1]　Chika Gonja[2]　Kaoru Kobayashi[3]

1. Introduction

Although introduced in the very beginning level of the English language program, acquiring the use of personal pronouns is often difficult for students, even college students, in Japan. One of the simplest reasons is the obvious syntactic difference between Japanese, a pro-drop language, and English which requires a syntactic subject of a sentence. In addition, while English has a finite number of pronouns, Japanese language has various forms of pronouns, which makes learning English or other European languages difficult. For example, English "I" can be translated into various forms in Japanese such as *watashi*, *atashi*, *watakushi*, *ore*, etc., and the use is determined pragmatically. Thus, the concept of personal pronoun is different between English and Japanese. We will first review some theoretical explanations, and then introduce empirical research findings before describing our present research.

When a group of sentences in any language is considered a unified text, it contains cohesive resources that create texture (Halliday and Hassan, 1976). Halliday and Hassan (1976) classified these cohesive resources into four categories: reference, ellipsis, substitution, and conjunction. Pronouns play an important role in achieving reference in English, while in Japanese, reference can also be realized by zero-pronouns or omission of pronouns. This difference makes it difficult for Japanese to learn the usage of pronouns as anaphora, a cohesive resource (Izumi et al., 2005).

Many studies have found that Japanese students are not using English pronouns properly. Hagino (1991) found Japanese university students' heavy use of pronouns in English, and argued that it can be attributed to the interlanguage that Japanese learners of English (JLE) had developed in the process of acquiring English pronouns. Overusing pronouns seems to be a common problem among JLE regardless of their English proficiency level. Matsuda et al. (2020) analyzed overused words in English essays written by college-level JLE and found that both intermediate and proficient learners overused personal pronouns. Kobayashi (2010) focused his research on the acquisition of first-person pronouns among JLE by comparing their use by junior high school students, high school students and university students. He found that junior high school students use the first-person singular pronouns *I* and *my* most frequently. The use gradually decreases as students proceed to high school and to university while the use of the first-person plural "we-pronouns" increases as they proceed to university. When compared to native speakers (NS), college-level JLE use "I-pronouns" and "we-pronouns" more often, especially "I-pronouns." Many college-level JLE use *I think* as a sentence starter

[1] 生命科学部言語科学研究室　[2] マラヤ大学大学院　[3] 東京農業大学生命科学部
This work was supported by JSPS KAKENHI Grant Number JP20K00757.

(Ishikawa, 2011), and college-level JLE who are in remedial classes have difficulty acquiring cases of pronouns (Oshiro, 2012).

However, acquisition of the proper use of pronouns is crucial when writing any type of English prose as overuse or misuse of pronouns negatively affects the evaluation of the text. Yang and Sun (2012) studied argumentative essays written by Chinese EFL learners and found that misuse of pronouns (including overuse) was observed among lower proficiency EFL learners, which may have contributed to their lower ratings.

These studies all demonstrate the importance of acquiring correct use of pronouns. In this study, on the assumption that the use of pronouns in English textbooks affects learners' command of pronouns, we investigated the occurrences of pronouns in MEXT[1]-approved high school English textbooks and compared them with those found in written and spoken English corpora.

We created a corpus of English textbooks (the TB corpus) and analyzed it with two large-scale balanced corpora of written and spoken English: British National Corpus (hereafter, BNC) and Corpus of Contemporary American English (hereafter, COCA). We attempted to examine both the overall use of pronouns and the use of third-person singular and plural pronouns as these are not directly translatable into Japanese. We were particularly interested in how the recent revision of the *Course of Study* (hereafter, *C of Study*) has influenced the use of pronouns. The new *C of Study* has a greater emphasis on communication in academic settings. One of its clear objectives is to prepare students to be able to understand the main points, details, purposes of the information, and ideas about daily or social topics in order to use them to communicate with others (MEXT, 2018a). More specifically, it aims to foster knowledge and skills, the ability to think, the ability to make decisions, the ability to express oneself, the ability to learn, and to have integrity (MEXT, 2018b).

Here are our research questions:

1. Is there any bias in the pronouns used in textbooks?
2. Can a gender gap be observed in third-person pronouns?
3. Has the new *Course of Study* influenced the use of pronouns in textbooks?

2. Materials and methods

A total of thirty-five MEXT-approved textbooks of *English Communication I* from twelve publishers published in academic year 2022 were used for the study. These textbooks are used in the subject *English Communication I* which is a required course for high school first-year students in Japan. All high school students are supposed to use one of these textbooks, which means the textbook they use could influence their attitudes toward the English language. Among the 35 textbooks, 11 are based on the previous *C of Study* and 24 are based on the new version of *the C of Study* which was announced in 2018. Since the transition is now taking place, textbooks based on both versions of the national curriculum were available in 2022. The

[1] Ministry of Education, Culture, Sports, Science and Technology, Japan

textbooks are basically selected by the board of education in each prefecture, and they determine which textbook is used in each high school. Therefore, the adoption rate of each textbook differs, which makes some textbooks more widely used than others. In the present study, we compiled a corpus based on all the textbooks now in use to achieve a holistic view.

2.1. Corpora

Three corpora were used in this study: High school English language textbook corpus (the TB corpus), BNC and COCA, which consists of seven sub-corpora. Table 1 summarizes the TB corpus. Because the number of textbooks based on the old *C of Study* has decreased over the years, we were only able to obtain 11 textbooks that are currently used in high schools. For the new *C of Study*, there were 24 textbooks. Some lessons in textbooks overlapped with the older version of the same textbook, so it is expected that same expressions can be found in both sub-corpora.

Table 1. High school English language textbook corpus (TB corpus)

Courses of Study	Tokens	Types	Texts	Textbooks
Old (2009)	51,141	5,180	150	11
New (2018)	151,405	9,642	314	24

BNC and COCA are two well-known large-size corpora. As the title suggests, BNC contains language samples mainly from British sources, and COCA contains American English samples. We decided to use them because both corpora are publicly available, and they were designed to include balanced data.

Table 2. British National Corpus (BNC) and Corpus of Contemporary American English (COCA)

Corpora	Sub-corpora	Texts	Tokens
BNC		4,054	110,691,482
COCA	Academic	26,137	120,988,361
	Newspapers	90,243	122,958,016
	Magazines	86,292	127,352,030
	Web (general)	88,989	129,899,427
	Web (blog)	98,748	125,496,216
	Fiction	25,992	119,505,305
	Spoken	44,803	127,396,932
	TV Movies	23,975	129,293,467

2.2. Methods

Since personal pronouns are thought to reflect stylistic features of the text, we first compared the frequency of personal pronouns in the TB corpus with those in the BNC and COCA sub-corpora. First, word frequency data from all corpora were compiled to ascertain which sub-corpus in COCA resembled the consistency of personal pronouns in the TB corpus.

For the BNC and COCA sub-corpora, we searched for all personal pronouns and obtained the data on the web (Davies, 2004, 2008-). For the TB corpus we used the word list and concordance functions of

WordSmith Tools v.8.0 (Scott, 2022). To compare data of different sizes, the frequency data were converted to frequency per million. After that, all the pronouns we looked at were grouped into six groups, "I-pronouns," "you-pronouns," "we-pronouns," "he-pronouns," "she-pronouns," and "they-pronouns." Each group includes subjective, objective, possessive, and reflexive pronouns in addition to those used as possessive determiners.

After comparing the overall occurrences of pronoun groups found in the TB corpus with those from other corpora, we compared the occurrences of pronoun groups in the textbooks of the previous *C of Study* (the Old *C of Study* sub-corpus) and those in the current *C of Study* (the New *C of Study* sub-corpus). We were particularly interested in the total uses of "he-pronouns" and "she-pronouns" as they are used comparably in grammar, but pragmatically they can also be an indicator of the gender gap. For the comparison of the ratio of pronouns in corpora, a Chi-squared test was used.

3. Results

Table 3 summarizes the occurrences of personal pronoun groups from BNC, COCA and TB corpora. The frequencies were standardized (occurrences per million). It clearly demonstrates that total frequencies and the ratio of personal pronouns vary depending on the type of discourse. While the Academic sub-corpus in COCA contains the lowest number of personal pronouns, the TV Movies sub-corpus contains many more personal pronouns than any other corpora we looked at. The ratio of different pronoun groups is also strikingly different among the corpora (Table 3; Figure 1).

Table 3. Occurrences of pronoun groups in corpora (per million words)

Corpus	Sub-corpora	**Total**	he	she	I	you	we	they
BNC	Spoken & Written	**54,947**	13,104	7,266	11,958	8,486	5,213	8,920
COCA	Academic	**22,381**	4,403	1,758	3,331	1,282	3,824	7,783
	Newspapers	**42,681**	13,142	4,550	7,698	4,034	5,062	8,195
	Magazines	**47,760**	10,273	4,582	10,360	8,641	5,734	8,170
	Web(general)	**58,407**	10,057	3,507	16,831	11,699	7,095	9,218
	Web(blog)	**64,815**	7,971	2,931	22,023	13,413	8,381	10,095
	Fiction	**106,300**	28,070	21,707	28,526	12,648	6,092	9,258
	Spoken	**85,596**	12,401	4,903	23,022	20,132	13,577	11,561
	TV Movies	**139,377**	12,402	6,680	54,448	46,103	13,103	6,642
Textbooks (TB)	New *C of Study*	**74,819**	13,414	7,913	14,194	13,368	9,577	16,353
	Old *C of Study*	**82,243**	16,112	8,213	18,009	12,260	10,950	16,699
	Textbooks All	**76,718**	14,096	7,988	15,157	13,098	9,924	16,456

As the results show more pronouns were used in spoken discourse, e.g., the Fiction sub-corpus, the Spoken sub-corpus and the TV Movies sub-corpus in COCA. Written corpora vary depending on the media and genre and generally have fewer numbers of pronouns. The TB corpus has comparatively more occurrences of personal pronouns for a written corpus. Comparing the New *C of Study* corpus and the Old *C of Study* corpus, the total frequencies of personal pronouns have decreased. From this analysis, the use of "I-pronouns," "we-pronouns" and "he-pronouns" has decreased, but "she-pronouns" and "you-pronouns" increased. The use of "they-pronouns" is more frequent in the TB corpus than any other corpora, and the ratio of "they-pronouns" in the corpus is also high, just like the Academic sub-corpus (Table 3). Unlike other pronouns in this study, "they-pronouns" refer to both animate beings such as people and animals and also inanimate objects, so they can be used in all four discourse types: narrative, argumentative, descriptive and expository. While narrative discourse may contain more personal pronouns, in academic writing which generally requires objectivity, personal pronouns are less used.

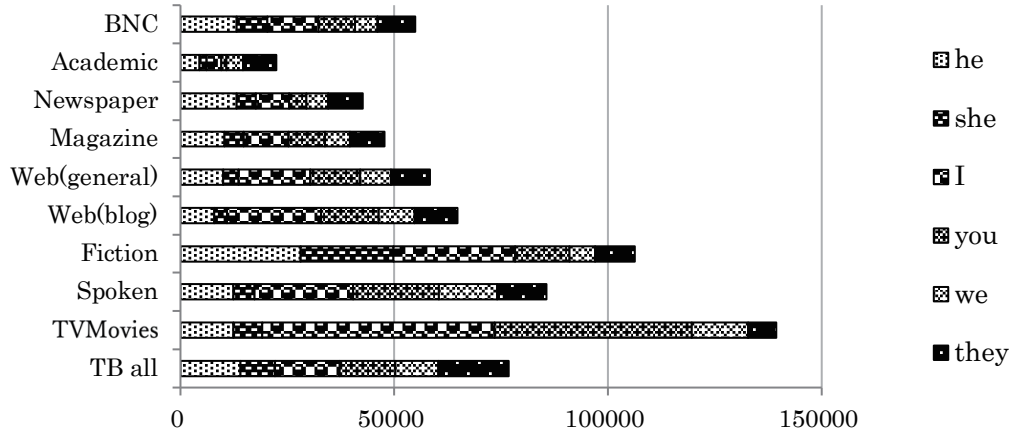

Figure 1. Comparison of the use of personal pronouns (c.f., Table 3)

Table 4. Occurrences of pronoun groups in the TB corpus (frequencies)

	he	she	I	you	we	they
Old *C of Study*	1069	606	1005	627	560	854
New *C of Study*	2660	1738	2397	2024	1450	2476
Multiple comparison	2.539*	-1.378	3.676**	-4.252**	0.892	-1.999*
	-2.539*	1.378	-3.676**	4.252**	-0.892	1.999*

$\chi^2(5)=36.873$, $p<.01$ Cramer's $V = 0.046$ *$p<.05$ **$p<.01$ (Benjamini-Hochberg)

Table 4 shows the actual frequency of the TB corpus. The ratio of each pronoun group was compared using a Chi-squared test. The results show that “he-pronouns” and “I-pronouns” decreased, and “you-pronouns” and “they-pronouns” increased. From this result and the result from Table 3, it is evident that “you-pronouns” are more frequently used in textbooks under the new *C of Study,* and the occurrences of “I-pronouns” and “he-pronouns” significantly decreased. Between “he-pronouns” and “she-pronouns,” the number of “she-pronouns” per million increased, but the increase is not significant.

4. Discussion

The results we obtained clearly show the distribution of personal pronouns can reflect the characteristics of the genre and discourse type. Having more personal pronouns may indicate that the corpus is spoken or narrative. On the other hand, academic texts have limited use of personal pronouns, even though the academic sub-corpus most likely contains texts from different subject areas. Given all these characteristics in different corpora, we will attempt to analyze the TB corpus. First, we question whether the language used in the textbooks truly represents the target variety of the language students need to acquire in high school. After that, we discuss if the gender gap is present in high school textbooks. Finally, differences between the old *C of Study* textbooks and new *C of Study* textbooks will be mentioned.

One of the characteristics of the TB corpus is the frequent use of personal pronouns. In our study, we found less authentic spoken language such as in TV programs and movies (the TV Movies sub-corpus) contain significantly more personal pronouns than natural discourse (the Spoken sub-corpus), which suggests the TB corpus, which contains comparatively more personal pronouns, may be simulated naturalistic discourse with some emphasis on personal pronouns. This can be explained as a sort of “teacher talk” variety of English, but since personal pronouns in English are quite different from those in Japanese, the textbook examples may result in the acquisition of overused pronouns as English input outside the classroom is limited for most students. Newer textbooks have fewer uses of personal pronouns, but their overall frequency is still higher than in the written language corpus or the BNC, which is a balanced corpus of English. This result could be one of the reasons why Japanese college students overuse “I” in their essays (Ishikawa, 2011; Kobayashi, 2010; Matsuda et al., 2020).

Another overused example is “they-pronouns.” “They” can refer to both people and objects, so even in academic English, they are used quite often. However, the occurrences of “they-pronouns” in the TB corpus are much more frequent than in any other corpora we looked at. Between the old and the new *C of Study* sub-corpora, the number per million is slightly lower, but the ratio is not. Other plural pronouns are also used frequently in the TB corpus, so it may be a general tendency of the TB corpus. This, again, may yield some thoughts about pronoun acquisition. As we do not have exact equivalents of those pronouns in Japanese, it is understandable for the textbook writers to include more examples of the target vocabulary, but they must also be aware that students also learn the pragmatics or the discourse styles of the language as well. Overusing particular pronouns in textbooks should be avoided. Since MEXT puts emphasis on academic communication competence in English, students are expected to be exposed to discourse similar to authentic

or natural academic discourse. If the textbooks deviate from the natural use of pronouns, it will make it difficult for students to learn how to properly use them and could explain students' awkward ways of using them in both speaking and writing.

Another important issue is the gap between female and male pronouns. Unlike other pronouns, the use of these two can be the same if both genders are treated equally in textbooks. However, our results show more uses of "he-pronouns" and less use of "she-pronouns." In fact, "she-pronouns" were fewer in all corpora we examined. If textbooks want to teach a simulated English language variety rather than completely authentic English discourse, then the writers of the textbooks should be more aware of gender-differences as students of a future generation need to be aware of a gender-gap free society. The imbalance of the overall usage of "she-" and "he-pronouns" in this study corresponds to what Ruddick (2010) argued in his gender analysis of an English language textbook used in a Japanese high school. In an attempt to achieve gender equality, main learning materials such as MEXT-approved textbooks must be free of the gender gap. In order to verify this, we need to add more data and study the actual contents of the lessons of each textbook.

Comparing the old and new textbook sub-corpora, it was observed that the number of "you-pronouns" increased in the textbooks under the new *C of Study*. This is thought to reflect the policy of strengthening the ability to communicate with others in the new *C of Study* (MEXT, 2018a). Strengthening communication skills has been an issue in English education in high school education, but there is a view that the old *C of Study* (MEXT, 2018b) was insufficient in strengthening the ability to exchange opinions with others. Although "I-pronouns" and "he-pronouns" have decreased, they are still used more than in BNC contributing to the heavy use of personal pronouns in the TB corpus as we discussed earlier. Further analyses of the contents are needed to clarify why personal pronouns are overused in textbooks.

In this study, we compared frequencies of pronouns in the TB corpus with BNC and COCA. We found that the TB corpus has more personal pronouns and "they-pronouns" for a written corpus. This is problematic because textbook texts must exemplify the use of pronouns in line with language used authentically. Another problem was that among the personal pronouns used in the textbooks, "he-pronouns" were used more than "she-pronouns." There should be no bias in the gender used in textbooks as students are to enrich their personality through English education (MEXT, 2018b). Finally, an increase in "you-pronouns" was observed in the new *C of Study* textbooks. This is in line with one of the policies for new *C of Study* which is to improve students' communication skills. Further studies on the contents of the texts in each textbook are needed to verify the above findings.

References

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Davies, Mark. (2004) *British National Corpus* (from Oxford University Press). Available online at https://www.english-corpora.org/bnc/.

萩野博子. (1991).「英作文での代名詞（3 人称）の使用法の考察」『日本実用英語学会論叢』 1991(1), 50-62.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. English Language Series, London: Longman.

Ishikawa, S. (2011). Phraseology overused and underused by Japanese learners of English. In K. Yagi, T. Kanzaki, & A. Inoue (Eds.), Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan (pp. 82-93). Kwansei Gakuin University Press.

Izumi, E., Uchimoto, K. & Isahara, H. (2005). Error annotation for corpus of Japanese learner English. In Proceedings of the sixth international workshop on linguistically interpreted corpora (LINC-2005).

小林雄一郎. (2010) 「日本人学習者の英作文における人称代名詞について」『言語処理学会第 16 回年次大会発表論文集』, 1074-1077.

松田紀子, 石井隆之, 岩田雅彦, 西美都子, 濱崎佳子. (2020).「英語学習者のエッセイに見られる過剰使用語―学習者コーパスの構築を視野に入れて―」『近畿大学総合社会学部紀要』 *8*(2), 19-27.

文部科学省. (2018a). 『高等学校学習指導要領』 Retrieved from https://www.mext.go.jp/content/1384661_6_1_3.pdf

文部科学省. (2018b). 『高等学校学習指導要領解説 外国語編・英語編』 Retrieved from https://www.mext.go.jp/content/1407073_09_1_2.pdf

大城明子. (2012). 「大学初年時基礎英語クラスにおける筆記小テストについて」『沖縄国際大学外国語研究』15(1), 中扉-1.

Ruddick, M. (2010). A Gender Analysis of An English Language Textbook Used in A Senior High School. 『新潟国際情報大学情報文化学部紀要』13, 11-29.

Scott, M. (2022). *WordSmith Tools version 8* (64 bit version) Stroud: Lexical Analysis Software.

Yang, W. & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and education, 23*(1), 31-48.