Unveiling Pronoun Patterns: Distinguishing Native English Speakers and Japanese Learners through Computational Linguistic Analysis

Abstract:

This paper investigates the pronoun usage of Japanese learners of English by analyzing data from the NICT Japanese Learner English (NICT_JLE) Corpus, which includes samples from both native English speakers and Japanese learners. Given that pronoun usage is more prevalent in English than in Japanese, this study explores how these linguistic differences manifest in Japanese learners' English proficiency. By applying TF-IDF vectorization, t-tests, and statistical analysis, we uncover systematic differences in pronoun patterns, with Japanese learners showing higher TF-IDF scores for pronouns compared to native speakers, indicative of a lower overall frequency of pronoun use. The study also deploys WordCloud and Bag-of-Words models to visualize these patterns, revealing that Japanese speakers rely heavily on pronouns like "you" while avoiding reflexive pronouns. This research contributes to a deeper understanding of second language production, with implications for Native Language Identification, language learning tool development, and NLP applications.

1. Introduction& Research Questions

This paper focuses on analyzing the pronoun usage of Japanese learners of English, based on the data from the NICT Japanese Learner English (NICT_JLE) Corpus. Pronoun usage is a crucial aspect of language proficiency, and understanding how learners navigate this feature can provide insights into both linguistic challenges and second language acquisition. Pronouns are used more

frequently in English than in Japanese, where speakers often rely on context or omit pronouns altogether. This cultural-linguistic difference often leads to notable patterns in the way Japanese learners use English pronouns, making it an important area for study in natural language processing (NLP).

The NICT_JLE Corpus is a collection of spoken English data from Japanese learners about the text file from the TOEFL test speaking part from both Japanese learners of English and native English speaker. This dataset is highly relevant to NLP because it captures authentic, spontaneous spoken language, offering valuable information for modeling non-native English speech. Furthermore, the corpus includes native English speaker data, allowing us to compare the two groups and identify systematic differences in language use. These comparisons are particularly useful for NLP tasks such as native language identification (NLI) and improving automated language learning tools by targeting specific areas of difficulty, like pronoun usage.

By comparing the pronouns usage of these learners with native English speakers from the same dataset, we aim to explore the following research questions:

1. How does pronouns usage differ between native English speakers and Japanese learners of English?
2. Which specific kinds of pronouns (e.g., first-person, third-person) are more or less frequently used by Japanese learners compared to native speakers?
3. Which pronouns (e.g., personal, possessive, reflexive) are over- or under-used by learners compared to native speakers?

The answers to these questions could have significant implications for improving language learning tools, advancing search in second language acquisition, and enhancing methods for identifying native languages based on spoken English patterns.We hope to contribute to a deeper

understanding of second language production, with potential applications in NLP tools designed to assist language learners.

2. Background

Research in second language acquisition has consistently shown that learners' native language (L1) exerts a strong influence on their second language (L2) learning. This is especially evident in the case of Japanese learners of English, whose native language allows for the omission of overt subjects, such as pronouns, in tensed clauses. Kuribara (2004) explored this issue in depth, finding that Japanese learners tend to omit pronouns due to L1 transfer. Japanese is a pro-drop language, meaning that speakers frequently leave out subjects (including pronouns) when they are understood from context. This structural difference leads to challenges when Japanese learners speak English, where subjects (particularly pronouns) must be expressed more explicitly. These findings support the theory that early Second Language Acquisition is heavily influenced by the learners' native language structure, resulting in systematic differences in pronoun usage between native and non-native speakers.

In the field of natural language processing, analyzing the language patterns of non-native speakers has become increasingly important, especially in the context of Native Language Identification, which aims to identify the speaker’s native language based on their L2 use. Several studies have applied machine learning techniques to this task. For example, Mayfield & Jones (2001) employed a Naive Bayes classifier to detect non-native utterances, focusing on lexical and syntactic features that distinguish native English speakers from Japanese learners. Their work provides insights into

feature selection for Native Language Identification tasks, highlighting how patterns in pronoun usage and sentence structure can reveal the influence of a speaker's native language.

The NICT Japanese Learner English (NICT_JLE) Corpus, which forms the basis of our study, is highly relevant for examining these issues. The corpus contains a large collection of transcribed speech data from Japanese learners, along with a comparison set from native English speakers, making it suitable for detailed analyses of grammatical and lexical features, including pronoun usage (NICT, n.d.). The dataset's rich annotation allows for the identification of disfluencies and patterns characteristic of non-native speech, which are important for both Second Language Acquisition research and NLP applications such as automatic error detection and learner language modeling.

More recent work has advanced the application of NLP techniques to Native Language Identification using more sophisticated models. Malmasi & Dras (2018) used classifier stacking and ensembles to improve Native Language Identification performance. Their approach utilized a variety of algorithms and configurations, with success using TF-IDF weighting of n-grams. The study demonstrated that lexical and syntactic markers, such as the frequency and usage of pronouns, are effective indicators of a speaker's native language.

By examining the pronoun usage of Japanese learners of English through the lens of these prior studies, we aim to build on this body of work. Our analysis, using NLP approaches such as TF-IDF and SVM will contribute to a better understanding of how non-native speakers navigate grammatical features like pronouns in their L2 production. This will not only further Second Language Acquisition research but also have potential applications in developing more effective language learning tools and improving Native Language Identification systems.

3. Data Collection and Preprocessing

In this part, we will introduce the characteristics of the corpus we used. We used tokenization and Sample balancing method to finish data preprocessing, and the reason why lemmanization and stopwords removal was not used is listed in the article.

3.1 Data Collection

The data we used are Japanese Native Speaker (L1 Japanese) text data and English Native Speaker (L1 English). The text data are recorded answers from TOEFL iBT. As English Speaker data was limited to 20 files and each file has different length and there are more than 100 files for Japanese files, we combined the English text and Japanese text separately. All sentence data is being imported as original text files and being stored in a dataframe. Then, tokenization work is being conducted.

3.2 Data Preprocessing

What needs to be noted in tokenization is that we did not perform stopwords removal and lemmatization because pronouns are contained in stopwords, if a removal is used, the data would not show the intended performance we need. The same with lemmatization: if lemmatization is used, all pronoun categories (for example, “he” and “his”) will be eliminated（all categories reduced to only “he”） and will mix the subject pronoun and object pronoun (“him” will be lemmatized as “he”).

There is already research in conducting TF-IDF and concatenating multiple text files into one when conducting preprocessing work (N, Zainuddin, 2014). As a result, we can obtain two data sets

incorporating both English and Japanese native speaker data with similar sentence unit amounts: 18565 units for Japanese samples and 5820 units for English samples. As there is less English data compared to the Japanese one, we used the method of oversample, and multiplied the English data by 3 times to make a comparison to the Japanese sample data. After combining the files, we performed text cleaning to clean the symbols other than text.

Also, Bag of Word Model (BoW Model) is used to show an insight of the pronoun preferences between English Native Speakers and Japanese Native Speakers.

4. Data Analysis

4.1 Text Vectorization & Feature Extraction

This section firstly introduces some of the data wrangling and analysis methods, and presents two visualizations and discusses why we used log transformation and t-test, and introduces how pronoun usages of pronouns in English L1 and Japanese L1 are different.

When Japanese L1 speakers speak English, there is a significant trend in using less pronouns compared with English L1 speakers. (KURIBARA, C. ,2004). The language usage feature stands for a Pro-Drop phenomenon that is widespread in the Japanese language. Pro-drop phenomenon means the first instance of a certain pronoun is used, the user of the language will not use the pronoun in the following context, given that the information receiver knows the subject of the language. Given the language features, Text Frequency- Inverse Term Frequency (TF-IDF) would be a suitable approach to solve the problem: it scores higher frequently instanced words with lower scores and lower instanced scores with higher scores (Sebastiani, F. 2002), and this rule could be applied in computing the weight for pronoun usages in text.

From the nature of TF-IDF, we expect there are lower scores on English Native Speakers generated pronouns, and higher on Japanese ones. Figure 1 recorded the visualization of TF-IDF score for Japanese Native Speaker text, and Figure 2 recorded visualization of English Native Speaker Text.

Comparing Figure 1. and Figure 2., it is easy to draw a conclusion that almost all of the pronouns, have difference in usages between Japanese native speakers and English native speakers: there are blanks in Japanese speakers in very low scores (0.0-0.2) , which means that the instances of those pronouns are relatively low. On the other side of the spectrum, the English native speaker data showed a different picture: they cluster in lower scores (0.0-0.2), which represents that there are relatively higher instances in the usage of those pronouns

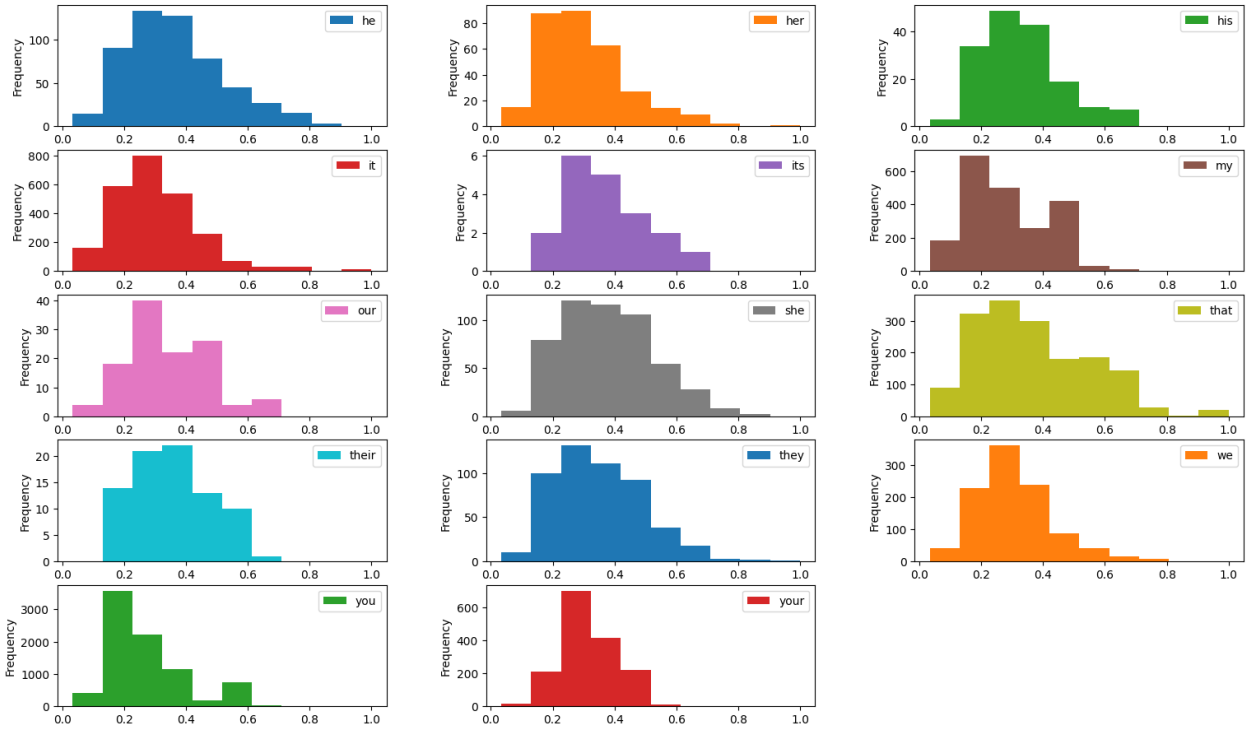

Figure 1. Japanese Initial TF-IDF Score

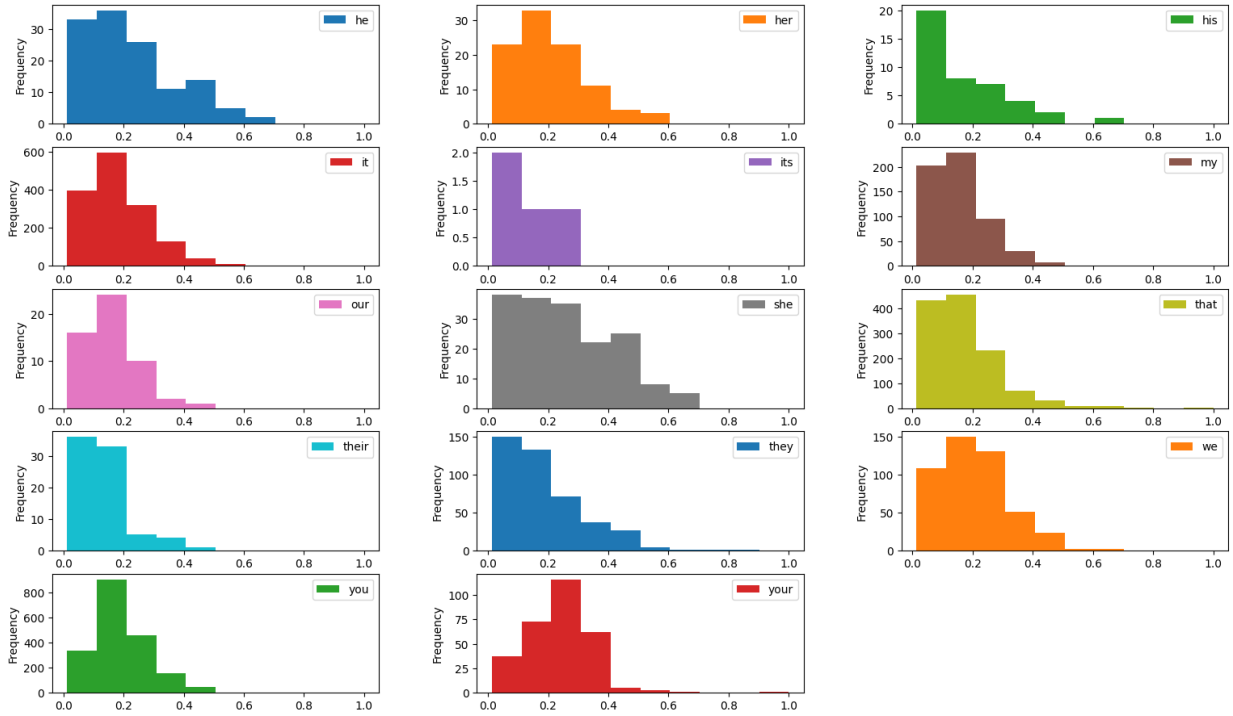

Figure 2. English Initial TF-IDF score

4.2 Mean Comparison

Figure 3 is a histogram that shows the differences of mean of TF-IDF score in Native English Speakers and Japanese learners of English to show the relative importance for each entry(Jiang, J., 2011). From the visualization we can see that all the pronouns being generated by Japanese speakers have apparently higher mean than English generated ones. The mean of the scores will be listed below in Table 1.

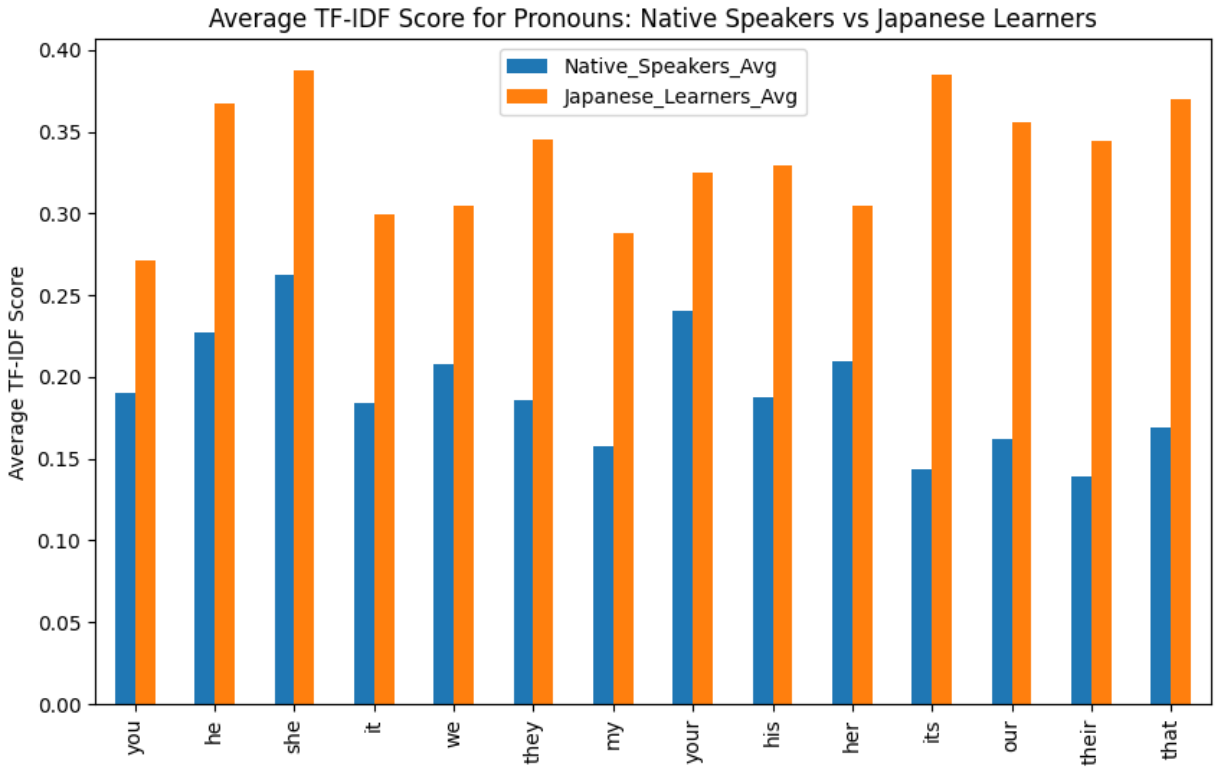

Figure 3.Visualization of Mean of English Native Speaker and Japanese Native Speaker Generated Pronoun TF-IDF Score by Pronouns

	Native_Speakers_Avg	Japanese_Learners_Avg
you	0.190206	0.271539
he	0.227152	0.367525
she	0.262792	0.387478
it	0.183695	0.299409
we	0.207697	0.304421
they	0.185808	0.345551
my	0.15774	0.288351
your	0.239999	0.32452
his	0.187264	0.328946
her	0.20966	0.304773
its	0.143811	0.384608
our	0.162277	0.355834
their	0.13887	0.34396
that	0.169283	0.369849

Table1. Mean of English Native Speaker and Japanese Native Speaker Generated Pronoun TF-IDF Score by Pronouns

One of the most commonly used statistical tests for hypothesis testing is t-test, however, it requires normal distributed data to make meaningful inferences (Ghasemi, A. et. al 2012). After Shapiro test, the data patterns show they are not in normal distribution ($p = 8.284657750934992e-216$ for Japanese data and $p = 1.522229899291352e-185$ for English), a log transformation is used to make them closer to normal distribution (Osborne, J., 2002). The visualization of the data is shown in

Figure 3. We now hypothesize that the null hypothesis is English and Japanese native speakers have the same usage of pronouns when speaking English.

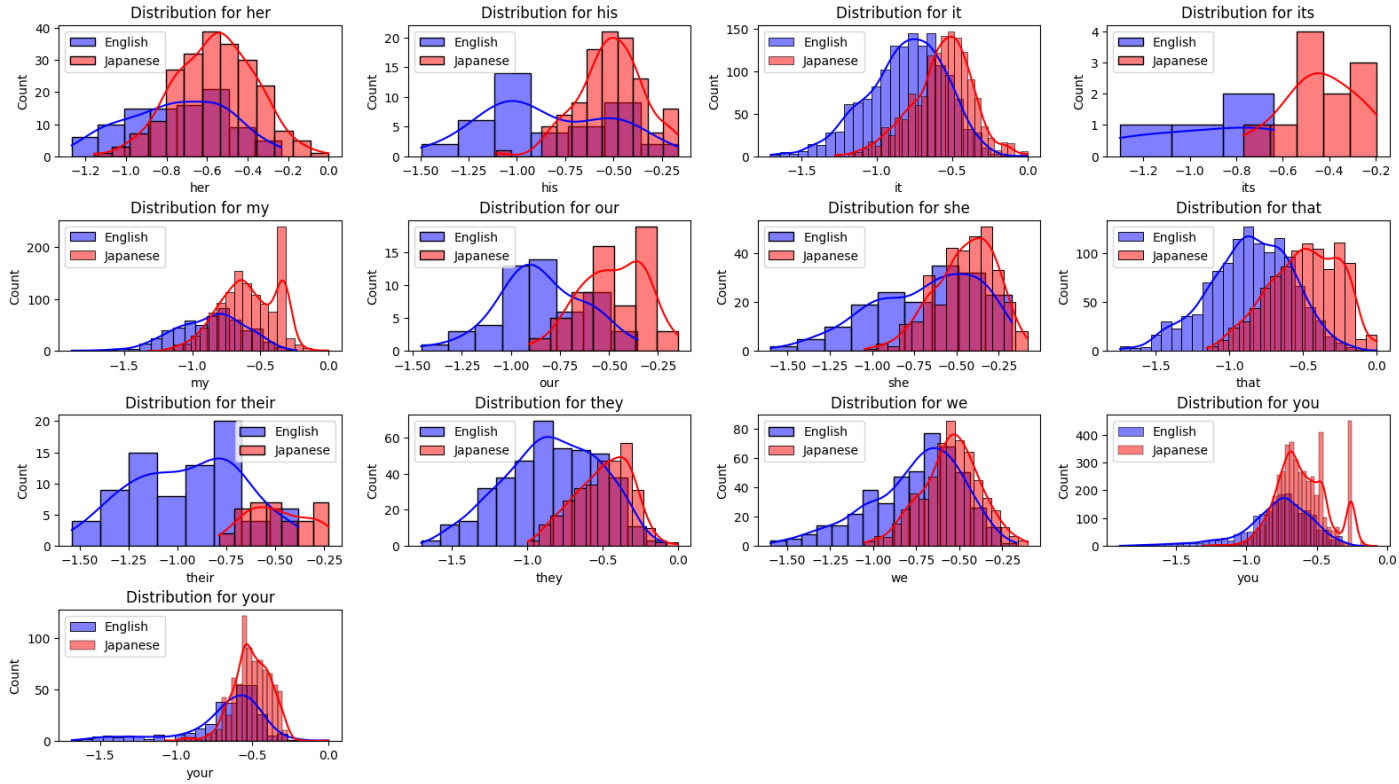

Figure 3. English & Japanese TF-IDF Score after log 10 transformation

It is still apparent that all of the pronouns could divide in a very apparent way. The t-test results are shown in Table 2.

Pronoun	P-value
he	1.16E-15
her	7.74E-10
his	7.70E-08
it	2.20E-153
its	4.86E-02
my	3.14E-91
our	4.83E-16
she	2.90E-16
that	4.97E-193
their	1.97E-19
they	6.64E-61
we	1.47E-41
you	9.47E-149

Table 2. pronoun t-test for log10 transformed English and Japanese Native Speaker Data

All of the result p-values are less than 0.05, and have a very small value. We have enough statistical evidence to reject the null hypothesis. Therefore, we can confirm that English and Japanese Native speakers do have significant differences in using pronouns when speaking English.

4.3 Pronoun Usage Frequencies

At last, as a supplement, BoW Model and WordCloud is deployed to show the preferences of the both groups of language speakers as BoW could effectively shows basic features of language like word usage preference (Tetreault, J. et. Al, 2013). The WordCloud showing the relative frequency of pronoun usages are listed. As Figure 4. shows English Native Speakers have a stronger tendency to use “it”, “you” and ‘That”, while they have relatively less tendency to use reflexive pronouns (myself, herself, himself, etc) in the given corpus. While for Japanese speakers, they tend to use “you”, “my”, “that”, “it” in the corpus, and they also have a very shortage in reflexive pronouns. Further discussion on the specific usage of the pronouns is needed.

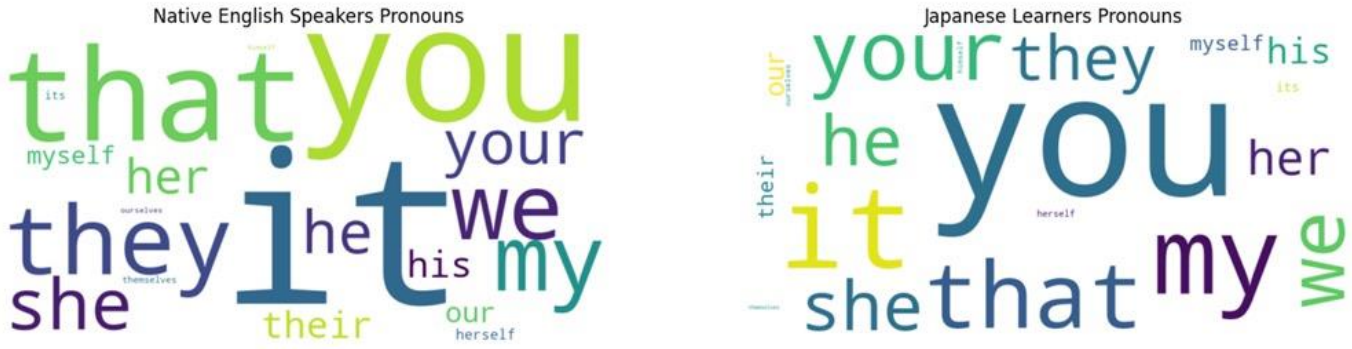

Figure 4. WordCloud for pronoun usage of English and Japanese Native Speakers

Figure 5. shows that English native speakers use “it” in the first place, and use “you” and “that” in the second and third place, while Japanese Speakers use “you” very frequently, while they use “it” and “my” in the second and third position, and the frequency drops dramatically.

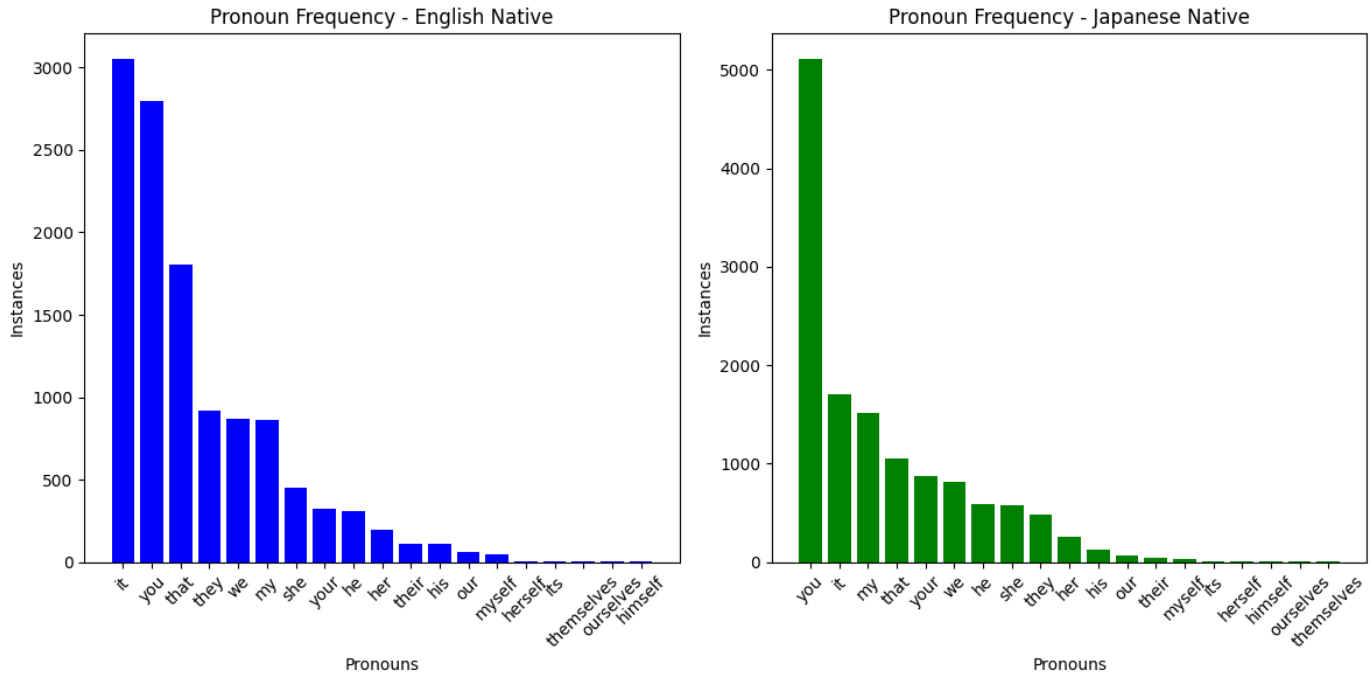

Figure 5. Histogram for the frequency of usage of Pronouns for English and Japanese Natives

4.4 Dimension Reduction and Modeling

This part of the article mainly explained how did the dimension reduction works to process with the TF-IDF data, how K-means works and why did I choose the combination of PCA and t-SNE for dimension reduction and visualization. What is more, Support Vector Machine (SVM) is

deployed for model training. The fitting rate result will be presented and further suggestions of improvement will be discussed.

In the dimension reduction step, I imported the TF-IDF matrix from American English L1 and Japanese L1, dropped labels, and used K-means for clustering—I set n cluster as 5 and random state as 42. At first, I assumed that clusters to be 3 – suggesting ordinary form of pronouns, processive form of pronouns and dummy subjects. However, after the visualization, it occurs to me that American English L1 speakers and Japanese native speakers do not stick to the pre-assumed pattern to use pronoun at all. Instead, they tend to cluster with more groups. Therefore, I set the clusters as 5. After setting the K-means clusters, I used Principal Component Analysis (PCA) for dimension reduction to 3 dimension to remain the principal feature of the matrix. Then, I used t-SNE to reduce the dimension to 2 for visualization. The visualizations are presented in Figure 6 and Figure7.

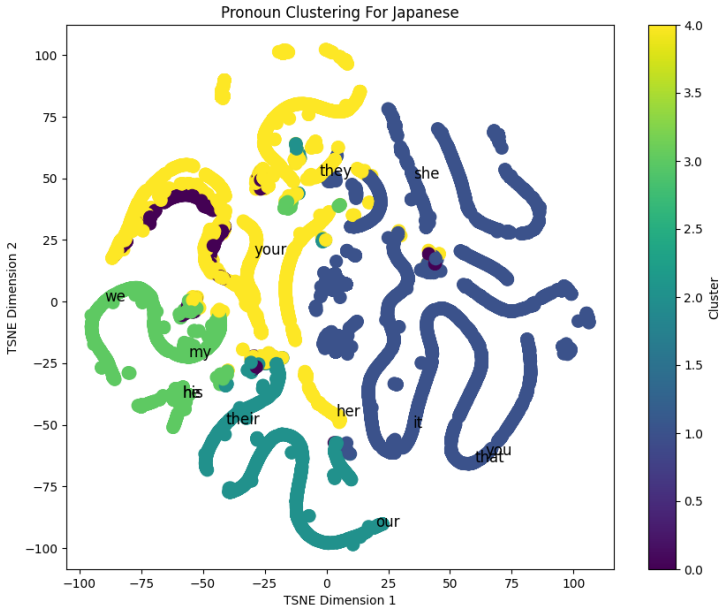

Figure 6. The t-SNE Pronoun Cluster for Japanese L1

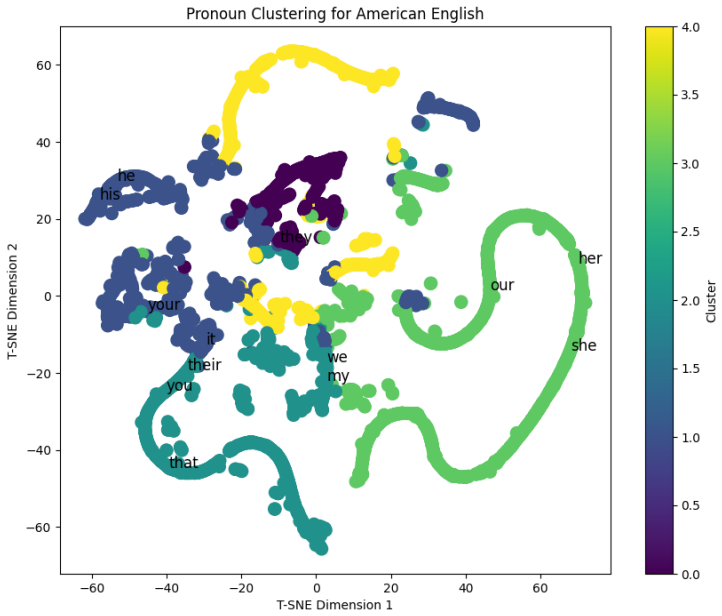

Figure 7. The t-SNE pronoun Cluster for American English L1

From Figure 6. and Figure 7. we can see that there are pattern differences between pronoun usages of Japanese L1 and American English L1: Japanese L1 speakers have clusters in “that”, “you”, and another cluster in “his”, “my”, “their”, “her”, “our “ . This means that there is a tendency in Japanese L1 speaker distinguishing the pattern of pronoun usage by pronoun forms: processive or ordinary form. Additionally, there is a strong tendency for Japanese L1 to cluster pronouns together, comparing with American English L1 speakers. While in English L1, clusters are mostly divided in pairs: “her” and “she”, “we” and “my, “ you” and “your” ---- this indicates that American English L1 know how to distinguish the usages of pronouns despite of the forms of pronouns. What is more, the pronoun behavior pattern for American English L1 is distributed in a relatively sparse pattern than the Japanese L1 speakers.

Support Vector Machine(SVM) is deployed in the modeling of the dimension reduced matrix. I labeled English L1 data as the index of 0 and Japanese as 1 and catenated the data into one dataset. 80% of the data are used for training and 20% of the data are used for examining the model. The results indicates that the accuracy rate is 0.81, which means there are 81% rate of correct prediction whether the sentence text is being generated by Japanese L1 speakers or being generated by American English L1 speakers. Therefore, we can firstly concludes that the model is effective and could perform a basic function is text classification.

5. Initial Inferences

The dataset's visual and statistical analysis of pronoun usage between native English and Japanese speakers reveals significant linguistic and cultural distinctions. Supported by strong T-statistics and P-values, these findings suggest that pronoun usage serves as a key feature for text classification based on language. The consistent separation seen in the data supports the hypothesis that distinct grammatical structures and cultural conventions heavily influence pronoun usage in each language.

The visualizations clearly show that English speakers use a broader range of third-person pronouns more frequently than Japanese speakers, who tend to omit pronouns in sentences, relying on context (zero anaphora). For example, in the case of "her" and "his," the T-statistics (-6.57 and -6.28) and extremely low P-values (7.74e-10 and 7.70e-8) demonstrate that English speakers use these pronouns significantly more than Japanese speakers. This aligns with the understanding that Japanese often omits possessive pronouns, particularly when the subject is understood within the context.

The most significant deviation is with the pronoun “it” (T-statistic: -28.11), which shows an even more dramatic difference between the two groups. In Japanese, “it” is frequently implied rather than explicitly stated, as subjects and objects are often dropped when they can be inferred from the context. This is particularly interesting as the avoidance of "it" in Japanese challenges direct translation methods and poses unique issues for machine learning models in natural language processing (NLP).

Cultural differences also seem to impact pronoun usage. The data reveals stark contrasts in the distributions for “we” and “you,” with significant T-statistics (-14.38 for “we” and -27.64 for “you”). In Japanese, the use of group-related pronouns may differ, as Japanese culture emphasizes collective identity and indirect speech, which might explain the lesser frequency of "we" and "you" compared to English. In contrast, English communication is more direct, requiring frequent use of explicit personal pronouns to maintain clarity.

The data for possessive pronouns such as “their,” “our,” and “my” also highlights these differences, with Japanese speakers using these significantly less than their English counterparts. This could reflect a preference in Japanese communication to omit such references when ownership or possession is understood in context.

Log transformations were applied to normalize the data and remove skewness, which further underscores the significant differences between English and Japanese pronoun usage. The distributions for pronouns like “our” and “their” become even more pronounced after

transformation, showing how normalization techniques reveal the underlying structure and patterns more clearly. The transformation provides a clear indication that the separation between the two linguistic groups is not just a result of raw frequency counts but is embedded in the linguistic norms of each language.

Given the significant statistical differences between English and Japanese speakers' pronoun usage, the dataset is highly suitable for text classification. The clear patterns in pronoun usage can be leveraged as features in machine learning models to classify texts based on linguistic origin. For instance, a classifier trained on these features could easily distinguish between English and Japanese texts, offering applications in language identification, educational tools, and more sophisticated NLP tasks, such as machine translation and pronoun resolution. The statistical and visual analyses strongly suggest that pronoun usage is a reliable linguistic marker between English and Japanese speakers. This makes the dataset an excellent candidate for classification models, which could further explore the impact of culture and grammar on language structure.

From TF-IDF Vectorization and t-test, a conclusion could be drawn: Japanese Native Speakers tend to use significantly less subject pronouns when speaking English compared with English native speakers. This phenomenon could be attributed to the pro-drop phenomenon, which means the language speaker tends to over pronoun when the prior context mentioned the subject.

Supplemental BoW Model and WordCloud Model were deployed on the corpus to show that there are higher instances for “it”, “you” and “that” for English Natives, and very high frequency on “you” but very low frequency on other pronouns for Japanese speakers. This could also be explained by a linguistic phenomenon called “dummy subject”, which means pronouns like “it” and “that” do not mean anything but leading the sentence. (Hussein, A. K, 2022). This phenomenon

is very common among the English language but the Japanese language does not have the concept. (Hirakawa, M. 2024) This infers that we not only need to detect the usages of grammar structures of pro-drop by Japanese L1 but also the usage of dummy subject by English Natives.

Go back to the research questions: 1. For the differences of pronoun usage patterns, the TF-IDF mean score and t-test shows, Japanese Learners of English tend to score higher in all pronouns than English speakers, which means the weight of pronoun in whole text is higher, the relative frequency of the pronouns is lower than English. 2. The WordCloud and BoW shows that Japanese Learners of English tend to use "you" while usage of others drop, while for English speakers, they tend to use "it", "you" and "that". This could be a reflection of the phenomenon of "dummy subject", which is unique to English and does not exist in Japanese. 3. Comparing the usage of different categories of pronouns is accessible, but due to the nature of the corpus itself, there are no reflexive pronouns being shown.

6. Further Inferences

The result of t-SNE dimension reduction shows that Japanese L1 clusters in a way that ordinary pronoun and processive pronouns are independent: "their", "my", "his", "her" are clustered together while plural forms of the pronouns, like "our", "we", "they" are outliers from the clusters. This shows the pronoun acquisition process of Japanese L1 speakers: in Japanese language, there is no natural form of pronoun and they need to catenate a certain pattern to accomplish the processive form of pronoun. What is more, the outliers of the plural forms provide evidence on the cultural of Japanese: Japanese L1 speakers mostly stress the importance of the collectives and community, and thus lead to the stress on the plural forms of pronouns. On the other side of the spectrum, American English L1 speakers shows a distinct pattern between different pronouns, and

they cluster the ordinary form and the processive form of the pronouns shows that they know how to distinguish the different usages of pronouns despite of the forms. What is more, there is a outlier on “that”, which is a sign of dummy subject usages. This could be interpreted as a sign of dummy subject structure while Japanese L1 does not obtain this feature.

The results on the t-SNE not only shows their differences on pronoun patterns but also shows that the differences on pronoun usages could be relative to the proficiency of the speakers. Further directions could focus on processing data from Japanese L1 with distinct English proficiency and comparing it with American English L1 speakers.

The result of SVM modeling shows the accuracy rate of 0.81, which is a effective prediction rate of accuracy. This could concludes that the distinguish English text generated by Japanese L1 and American English L1 through pronoun usage patterns could be an effective strategy and the model could be used for practical text classification scenario.

7. Limitations

There are still limitations on the research: The method of TF-IDF and BoW only focus on the usage of lexicon while ignores the syntactic patterns. To further the research, models that could detect syntactic patterns are needed. For instance, BERT can detect dummy subject, which marks the frequently use of “it’ and “that” (common in English natives) and pro-drop, which leads to less usage of pronouns overall (common for Japanese natives). The example of patterns will be presented in Figure 8 and Figure 9. Therefore, in the future work, we need to use BERT to find out the distribution of the pattern along with the weight and frequency of the certain pronouns by TF-IDF and SVM.

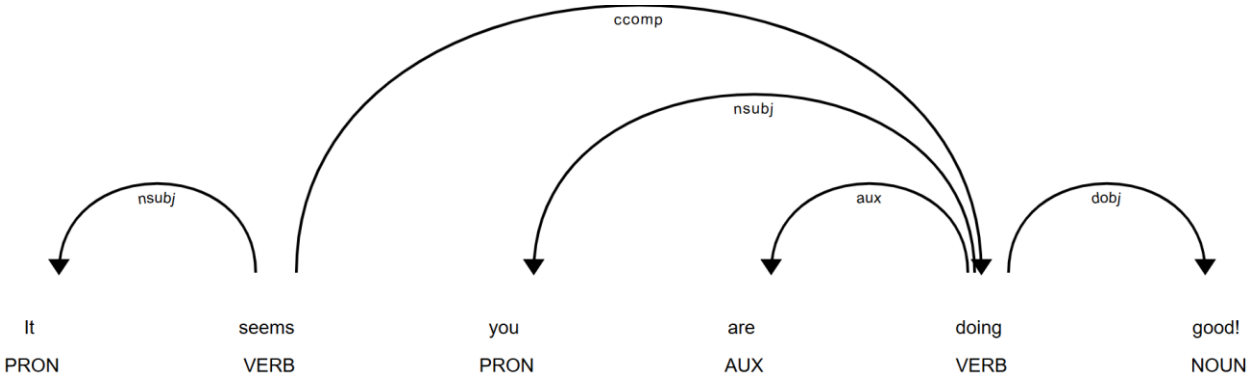

Figure 8. Example of Dummy Subject

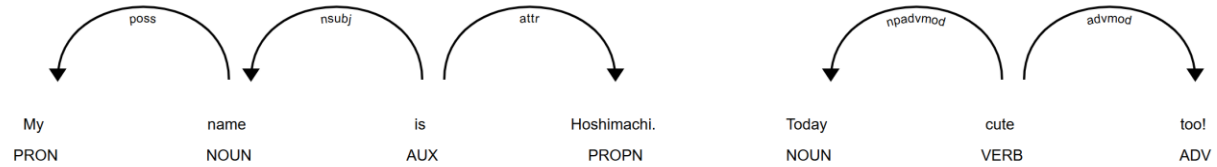

Figure 9. Example of Pro-Drop

8. Implications for Text Mining or NLP applications

The comparative analysis of pronoun usage between Japanese and English speakers offers valuable insights that have direct implications for text mining and NLP applications in cross-linguistic studies. One of the primary observations from this dataset is the significant structural and cultural differences in pronoun usage between these two languages. These differences challenge standard NLP approaches, especially when applying sentiment analysis, machine translation, or gender detection tools across languages.

For instance, in English, pronoun usage is relatively straightforward, and gendered pronouns (he/she) are explicitly tied to the subject's gender. However, Japanese pronouns are far more flexible and context-dependent, with a range of pronouns that vary based on politeness, formality, and social hierarchy. This complexity in pronoun selection reflects underlying cultural nuances, which are not easily captured by typical NLP models trained on more standardized datasets. This finding suggests that sentiment analysis and other NLP tasks may require specialized models that incorporate cultural and contextual understanding, as opposed to generic approaches that assume a universal mapping of pronouns to sentiment or gender.

The findings indicate that gender-related NLP tasks such as gender detection or sentiment analysis may encounter limitations in languages like Japanese, where gender-neutral pronouns (like "watashi" in Japanesesdd) are often used regardless of the speaker's gender. The lack of explicit gender markers in the pronouns challenges traditional gender classification models, which are typically dependent on linguistic cues such as pronouns or titles (e.g., "Mr." or "Ms." in English). This finding highlights the need for more sophisticated techniques in gender identification that go beyond just pronouns, incorporating additional context from the surrounding text or broader sociolinguistic features.

On a practical level, these insights suggest that cross-linguistic NLP applications should be designed with sensitivity to the unique syntactic and sociocultural factors that influence pronoun usage. For example, machine translation systems between Japanese and English may benefit from incorporating models that account for the cultural context of pronoun usage, potentially leading to

more accurate translations. Similarly, tools for gender detection or sentiment analysis in multilingual contexts will need to adapt to the varying ways pronouns are used, avoiding reliance on rigid, language-specific rules.

The dataset reveals how language-specific features, such as the varying levels of formality in pronoun use, could affect the interpretation of sentiment and politeness in dialogue systems. In Japanese, the choice of pronoun can signal respect or familiarity, influencing how sentiment is conveyed and understood. This underscores the importance of incorporating sociolinguistic factors into NLP models, particularly when developing tools for conversational AI or virtual assistants that need to navigate cultural nuances.

In summary, the insights drawn from the comparative analysis of Japanese and English pronoun usage point to the limitations of applying generic NLP models across languages without considering the unique linguistic and cultural characteristics that shape how language is used. This highlights the need for specialized approaches in text mining and NLP applications that can adapt to the complexities of multilingual, cross-cultural datasets.

References:

1. Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. International Journal of Endocrinology and Metabolism, 10(2), 486-489. https://doi.org/10.5812/ijem.3505

2. Hirakawa, M. (2024). Null subjects and long-distance anaphors revisited: What the acquisition of Japanese vs. Chinese contributes to generative approaches to SLA. Studies in Second Language Acquisition, 70, 1–30. https://doi.org/10.1075/lald.70.01hir

3. Hussein, A. K., & Al-Hussein, A. M. A. (2022). Dummy subjects in English: A grammatical analysis. International Journal of Linguistics, 14(4), 120-133. https://doi.org/10.5296/ijl.v14i4.20263

4. Jiang, J., & Khoshgoftaar, T. M. (2011). A survey of data mining techniques for software engineering. Journal of Software Engineering and Applications, 4(1), 64-76. https://doi.org/10.4236/jsea.2011.41008

5. KURIBARA, C. (2004). Mistaanalysis of subjects in Japanese-English interlanguage. Second Language, 3, 69-95. https://doi.org/10.11431/secondlanguage2002.3.0_69

6. Malmasi, S., & Dras, M. (2018). Native language identification with classifier stacking and ensembles. Computational Linguistics, 44(3), 403-446. https://doi.org/10.1162/COLI_a_00323

7. Mayfield, L., & Jones, R. (2001). You're not from 'round here, are you? Naive Bayes detection of non-native utterances. In Proceedings of the Second Meeting of the North

American Chapter of the Association for Computational Linguistics (pp. 31-37). https://aclanthology.org/N01-1031

8. Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research, and Evaluation, 8(6). https://doi.org/10.7275/4vng-7h71

9. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47. https://doi.org/10.1145/505282.505283

10. Tetreault, J., Blanchard, D., & Cahill, A. (2013). A report on the first native language identification shared task. *In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 48-57.

11. Vajjala, S., & Banerjee, S. (2017). A study of n-gram and embedding representations for native language identification. *In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 240-248). Association for Computational Linguistics.

12. NICT. (n.d.). NICT JLE corpus. *National Institute of Information and Communications Technology*. Retrieved from https://www.nict.go.jp/en

13. Ishikawa, S. (2015). Phraseology overused and underused by Japanese learners of English: A contrastive interlanguage analysis. *Asian EFL Journal*, *17*(3), 135-159. Retrieved from https://www.researchgate.net/publication/285754551_Phraseology_overused_and_underused_by_Japanese_learners_of_English_A_contrastive_interlanguage_analysis

14. Sperlich, D. (2021). Japanese learners of English and reflexive pronouns. *Intercultural studies,3,* 3-12. https://www.komatsu-u.ac.jp/common/images/bulletin202103-1.pdf

15. Natsukari,S. (2013). Use of I in Essays by Japanese EFL Learners. *JALT Journal,61.* *https://doi.org/10.37546/JALTJJ35.1-4*