Usage of English Articles among L2 Learners
A corpus study of graded essays from TOEFL

Holly Zheng
yuqi_zheng@brown.edu

Linghai Liu
linghai_liu@brown.edu

Danielle Springer
danielle_springer@brown.edu

1. Introduction

English encodes definiteness with the definite article “the”, and noun phrases with ”the” usually refer to entities that are unique in the discourse context (i.e. unique definite) or entities that have been mentioned in previous discourse (i.e. anaphoric definite). However, many languages encode definiteness in ways other than using overt definite markers. Learning the definite article system in English, then, is a challenge for learners with native languages that handle definiteness in a different way.

In Mandarin, definiteness is encoded in two ways: unique definites can be encoded with a bare noun, and anaphoric definites are usually encoded with demonstratives. Indefinite noun phrases are presented as bare noun phrases or co-occur with numeral classifiers, and Mandarin lacks a counterpart of the English indefinite articles. [5]

In many Romance languages, such as Spanish, an explicit definite article does exist. One additional distinction between Spanish and English noun phrases is that noun phrases have gender markers in Spanish.

1.1. Problem Statement

In this project, we plan to look into how learners of English from different L1 backgrounds acquire and express definiteness in English. We focus on English texts produced by native speakers of Mandarin and Spanish, which are two languages that exhibit quite different definiteness marking systems. We utilize a corpus of graded essays from the Test of English as a Foreign Language (TOEFL), administered by Education Testing Service. We expect to investigate the following 2 hypotheses:

- Speakers of languages that lack an explicit definite article use their L1 language’s counterpart(s) of the overt English definite marker more frequently than speakers of native languages that have an explicit definite article. Specifically, we hypothesize that native speakers of Mandarin use demonstratives to encode definiteness more frequently than native speakers of Spanish using demonstratives.
- The distribution of uses between demonstratives vs. definite articles in English among Mandarin native speakers changes as the Mandarin speaker’s degree of advancement in English increases.

We would also include discussions on how the acquisition of the definite article system interacts with other factors during L2 learner’s timeline of learning English. Specifically, we will look at:

- How L2 learners of English with Mandarin L1 resolve indefiniteness, and whether the learning of the indefiniteness system fluctuates with using “the” in their writing. Our hypothesis is that L2 English learners would start by reducing the usage of indefiniteness and neutralizing them by one or singular demonstratives before using only the “a/an” articles.

1.2. Overview

Section 2 presents a review of relevant literature on L2 acquisition of the English definite article system and theories that will serve as references for our project. Section 3 explains the corpus and key data preprocessing steps. Section 4 presents our initial observations and statistical analysis based on a few models. Section 5 concludes the paper and presents potential future works. Section **??** contains our response to questions asked by students based on our presentation.

2. Related Works

There has been a wide breadth of work analyzing the linguistic challenges associated with L2 learning, and more specifically, L2 learning of English. Vajjala (2018) [9] asserts that one’s particular native language is an appropriate predictor of a TOEFL test taker’s score, some L1s being facilitating L2 acquisition, and others creating more challenges in L2 learning.

Other studies investigate particular challenges in the acquisition of the English definiteness system in L2 learners. Ionin et al. (2004) [4] assert that speakers with no L1 article system struggle to master the English definiteness system as a result of the Article Choice Parameter (ACP), the ability of an L2 learner to understand articles in a two-article language (such as English or Samoan) as either being characterized with the feature [specific] or [definite]. English, according to this paper, has no particular article which communicates [specific] but categorizes ”the” as [+definite] and ”a/an” as [-definite]. What follows is the Fluctuation Hypothesis, in which L2 learners of two-article languages will, at times, divide L2 articles into a definiteness system and, at other times, a specificity system. With enough L2 input, these learners may realize the correct system with which to categorize L2 articles. In English, the Fluctuation Hypothesis predicts L2 learners will make the mistake of overusing ”the” as an indefinite and overusing ”a” as a definite [4]. Tryzna et al. (2009) [8], by contrast, suggests an optionality of choice of articles. Ko et al. (2010) [6] examines data from adult Korean L2 English learners in order to analyze the role of presupposition in the learning of the English definiteness system, finding that the overuse of ”the” with indefiniteness may be the result of presuppositional indefinite contexts.

Another problem in L2 learning of the English definiteness system is the contrasting ways in which speakers process L1 and L2. Hawkins & Chan (1997) [2] support the Representational Deficit Hypothesis (RDH) and the failed functional features hypothesis, noting the inaccessibility of features of functional categories in L2 learning.

One of the challenges we encountered was finding the right proxy for determining the skill level of the TOEFL graded essays. Huttenlocher et al. (2007) [3] used proxies such as mean length of utterance, number of word tokens, and number of sentences, and number of unique words as measures of the complexity of speech. We referred to these metrics during the data preprocessing steps.

3. Data

The work presented by the relevant papers mainly focus on experimental approaches to investigate the acquisition of English articles in L2 learners. Many of the papers conducted studies where participants were asked questions about English grammar. In our current project, this experimental approach is not feasible, so instead we choose to analyze a corpus of natural text produced by L2 English learners.

In this project, we utilize a corpus of essays written by non-native speakers of English to investigate the distribution and usage of English definite articles in speakers of different native languages. Specifically, we focus on essays written by native speakers of Mandarin and Spanish. In this section, we will first introduce the corpus we are using and the preprocessing steps. Then, we will talk about the specific words we focus on and how we constructed proxies for an essay author’s proficiency in English.

3.1. Corpus

We utilize a corpus that contains graded essays from the Writing section of the TOEFL exam (Test of English as a Foreign Language), collected between 2006 - 2007. [1] TOEFL is a commonly used standardized test that determines non-native English speakers’ proficiency in English, and a high score on this test is often required during the application process to higher education institutions in English-speaking countries. The original corpus is released by the exam’s organizer, Education Testing Service (ETS), through the Linguistic Data Consortium (LDC). 11 different native languages are represented in the corpus, and each L1 language has 1,100 graded essays. The corpus contains tokenized essay responses as well as the score level each essay received on the exam, divided in to 3 categories: high, medium, and low.

Ideally, we would have utilized another corpus as the control group, and this control corpus would be graded essays by native English speakers from similar standardized tests. It was difficult for us to find a corpus that matches this criteria, however, so we chose to focus on analyzing differences exhibited by native speakers from different languages (Mandarin and Spanish, in this case), as well as differences between the score levels.

3.2. Initial Preprocessing

To preprocess the original corpus, we wrote scripts in Python to first extract the files that contain essays by Mandarin and Spanish native speakers, which is a total of 2,200 essays. For each essay, we ran the NLTK [7] tokenizer, lemmatizer, and Part-of-Speech tagger on each sentence to enable accurate calculations such as counting the number of nouns and the number of unique words in each essay. We calculated relevant information and compiled all data into one CSV file for each L1 language. Each file includes the following information:

- Score level as indicated by original corpus
- Number of words in the essay
- Number of sentences in the essay
- Number of unique words in the essay
- Total counts and frequencies of nouns, verbs, adjectives, and modal verbs (eg. ”do”, ”can”)
- Total counts and frequencies of the following words: the, a, an, one, this, that, these, those, the other words labeled as determiner by the POS tagger (eg. ”some”, ”every”, ”both”), and pronouns

3.3. Proxies for English Proficiency

Since the original corpus only presented 3 categories for score levels, we constructed a few of our own proxies for how advanced each essay author's English is. One possible proxy for the complexity of an author's English is the number of unique words in an essay. Intuitively, the more advanced an author is, the more varied vocabulary they should have. Also, essays that score higher tend to be longer in length, which can be reflected by a higher number of unique words as well. The number of unique word as a metric gives us a continuous rather than discrete measurement of English proficiency level. We plotted the density of the number of unique words for the Mandarin and Spanish corpus separately, and found that the increase in number of unique words does correspond to increase in score level, so we felt confident in using this metric as a proxy for an essay author's English proficiency. (Fig. 1, Fig. 2)

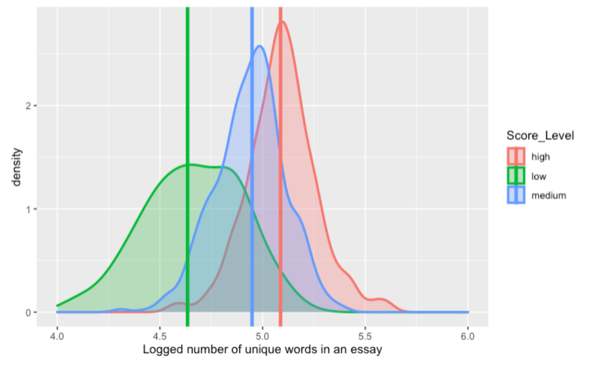


Figure 1: Logged number of unique words against score levels, L1 Mandarin

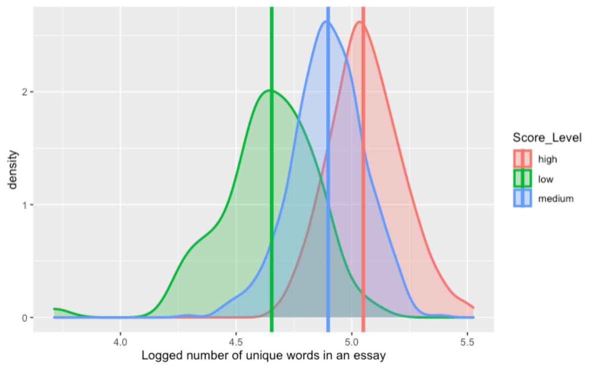
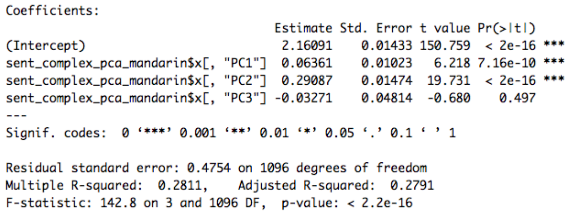


Figure 2: Logged number of unique words against score levels, L1 Spanish

Other potential proxies for English proficiency level include the average number of nouns per sentence and average sentence length. We performed a Principle Component Analysis on these three different proxies on English proficiency, with the essay's received score level as the predicted variable, in order to determine which proxy better predicts the essay's actual received score. Result of this PCA (Fig. 3, Fig. 4) shows that, although the data presents fairly high variance on score level and these proxies , the number of unique words contributes the most to predicting an essay's score level, so we will be using number of unique words as the proxy for English proficiency in this study.

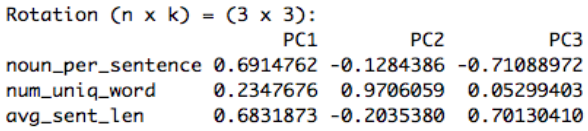
After establishing an appropriate proxy to transform categorical data of score level for the original corpus onto the set of real numbers (from discrete to continuous), it becomes plausible for us to experiment with regression models to observe and predict the distribution of language usage for L2 users of English (authors of our corpus), with respect to their English proficiency.



```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          2.16091    0.01433 150.759  < 2e-16 ***
sent_complex_pca_mandarin$x[, "PC1"]  0.06361    0.01023   6.218 7.16e-10 ***
sent_complex_pca_mandarin$x[, "PC2"]  0.29087    0.01474  19.731  < 2e-16 ***
sent_complex_pca_mandarin$x[, "PC3"] -0.03271    0.04814  -0.680    0.497    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4754 on 1096 degrees of freedom
Multiple R-squared:  0.2811,	Adjusted R-squared:  0.2791 
F-statistic: 142.8 on 3 and 1096 DF,  p-value: < 2.2e-16
```

Figure 3: PCA Coefficients on sentence complexity metrics



```
Rotation (n x k) = (3 x 3):
                        PC1        PC2         PC3
noun_per_sentence 0.6914762 -0.1284386 -0.71088972
num_uniq_word     0.2347676  0.9706059  0.05299403
avg_sent_len      0.6831873 -0.2035380  0.70130410
```

Figure 4: Principle Components for PCA on sentence complexity metrics

4. Analysis

4.1. Observations

Our hypotheses stated in problem statement are based on the assumption that the acquisition of English as a second language (L2) would be influenced by certain properties in the learner's native language (L1). Hence, we began by looking at the L1 Mandarin and L1 Spanish corpora separately and performing exploratory visualizations to understand the similarities and differences of each corpus' statistical trends.

Mandarin does not have explicit definite or indefinite articles, but the language does have demonstratives, possessive pronouns, and numerals. Definite and indefinite articles, demonstratives, possessive pronouns, and numerals are all present in Spanish as those are in English. The difference between the article systems of these two languages forms the basis for our overall hypothesis, which is that the acquisition of definite articles for L1 Mandarin speakers should be systematically different from that for L1 Spanish speakers.

We gathered information on the distribution of the frequencies of usage of various determiners in each corpus. In Table 1, we find that the variances of the distributions become smaller as score level increases for both L1 Mandarin and Spanish speakers, which could suggest that more advanced English speakers has a more unified behavior when using these articles. However, in the case of L1 Spanish, this pattern is more varied, so further nuanced investigation is needed for the two corpora.

We also provide observations to support the choice of variables for the modeling in the next section. Among definite and indefinite articles, demonstratives, possessive pronouns, other determiners, and the numeral 'one', we exclude the numeral 'one' in our analysis. More than 25% of essays in our corpus do not employ the numeral 'one'. The medians of the logged frequencies of "one" in both corpora indicate that, on average within each essay, there is only less than one occurrence of the word 'one'. (The mean essay length is 326.43 and 331.21 for L1 Mandarin and Spanish) Therefore, the effect of this numeral would be trivial in any modeling we plan to do. Even if we include it, we would be unable to explain more than 25% of the population. So, we decide to not investigate the usage of the numeral "one" in any following analysis.

4.2. Models and Analyses

4.2.1 "The" vs. Demonstratives

Our hypothesis is that, since demonstratives in Mandarin encode definiteness for the language instead of using an overt definite marker like "the", lower proficiency in English in Mandarin speakers would correspond a higher frequency of demonstrative usage, because beginner English learners might choose to use the the definite marker in their native language when writing in English. Since there exist overt definite articles in Spanish, such overuse of demonstratives should be reflected in the Spanish L1 corpus.

We ran linear regression models with number of unique words as the independent variable, and the log frequency of "the" and 4 demonstratives "this/that/these/those" as the dependent variables, while comparing results from both corpora. Fig. 5 shows that, for both word categories in the L1 Mandarin corpus, the data points of frequency are fairly scattered. The correlation for both word categories with respect to the author's proficiency in English is fairly weak. Fig. 6 shows a similar weak correlation for the Spanish corpus. We see a slight decrease in the frequency of demonstratives and a sight increase in frequency of "the" in Mandarin speakers as their English proficiency increases, but the comparison between these 2 plots shows that the relationship between English proficiency and the frequency of "the" and demonstratives is similar for Mandarin and Spanish L1 speakers.

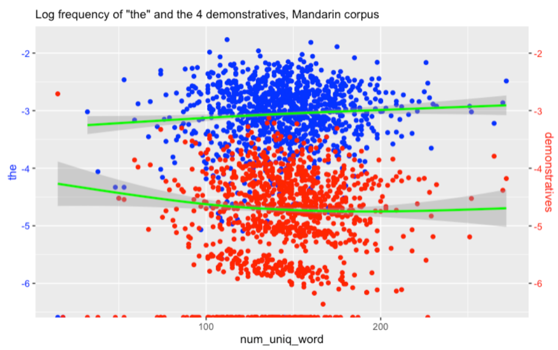


Figure 5: Log frequency of "the" and 4 demonstratives vs. number of unique words, L1 Mandarin

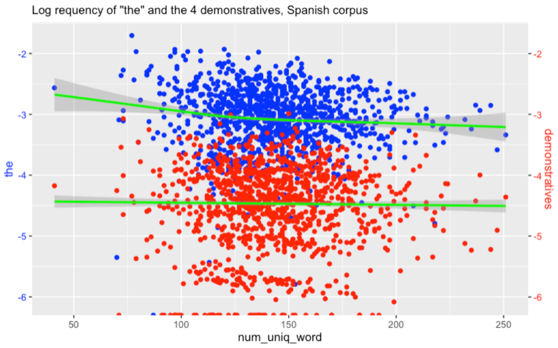


Figure 6: Log frequency of "the" and 4 demonstratives vs. number of unique words, L1 Spanish

To address our hypothesis regarding a change in use of demonstratives as Mandarin native speakers' proficiency increases, we further investigated the ratio between the logged frequencies of demonstratives and 'the' in both languages. As shown in Fig. 7, the distribution of the ratio is weakly correlated with our proxy for English proficiency.

Looking at density plots for the ratio between the number of "the" and the number of all demonstratives (Fig. 8), we find a slightly clearer pattern: we see an increase in the ratio between "the" and demonstratives as score level goes up for Mandarin native speakers. The medium and high score essays' density plots, though, mostly overlap, so we think that there is a slight overuse of demonstratives for beginner English learners from Mandarin L1 background. This interpretation should be taken with caution, as the regression results show weak correlation between these two variables. The density plot for Spanish L1 corpus (Fig. 9) shows that the ratio among all score levels is similar.

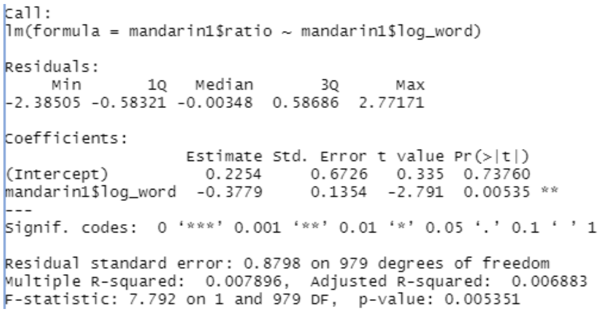
Table 1: Standard Deviation of Frequency of Usage, by Score Level

	DEF.art	IND.art	DEMO	PRP	Other.DET	ONE
Low	0.027348	0.014306	0.01118	0.01241	0.01067	0.006624
Medium	0.025944	0.014346	0.00716	0.01088	0.008002	0.005913
High	0.021570	0.01422	0.006524	0.009998	0.006463	0.005739

(a) L1 Mandarin

	DEF.art	IND.art	DEMO	PRP	Other.DET	ONE
Low	0.033852	0.016224	0.0096127	0.0086419	0.01004	0.0052507
Medium	0.025009	0.01384	0.0082232	0.0087153	0.0076155	0.0048778
High	0.020025	0.014049	0.0071127	0.0089451	0.0061835	0.0060469

(b) L1 Spanish



```
Call:
lm(formula = mandarin1$ratio ~ mandarin1$log_word)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.38505 -0.58321 -0.00348  0.58686  2.77171 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)   
(Intercept)          0.2254     0.6726   0.335  0.73760   
mandarin1$log_word  -0.3779     0.1354  -2.791  0.00535 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8798 on 979 degrees of freedom
Multiple R-squared:  0.007896,	Adjusted R-squared:  0.006883 
F-statistic: 7.792 on 1 and 979 DF,  p-value: 0.005351
```

Figure 7: ratio log $\frac{demo}{the}$ vs. log(unique words), L1 Mandarin

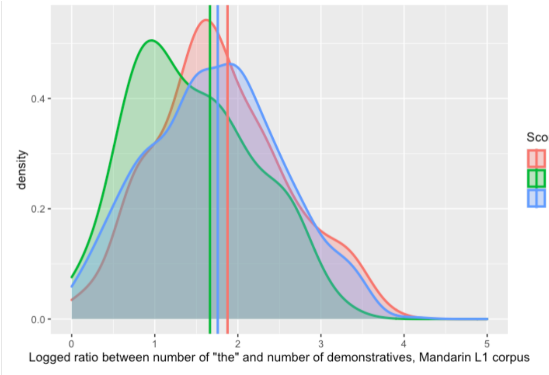


Figure 8: Logged ratio between ”the” and demonstratives, L1 Mandarin

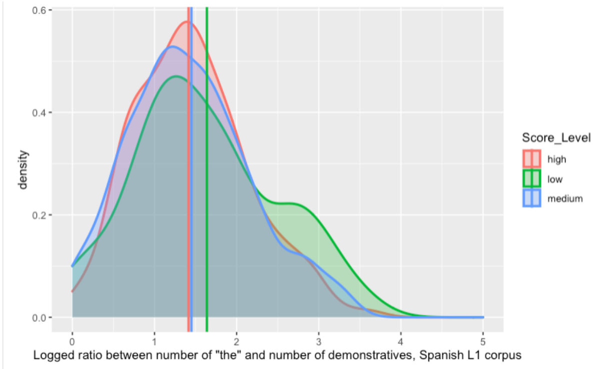


Figure 9: Logged ratio between ”the” and demonstratives, L1 Spanish

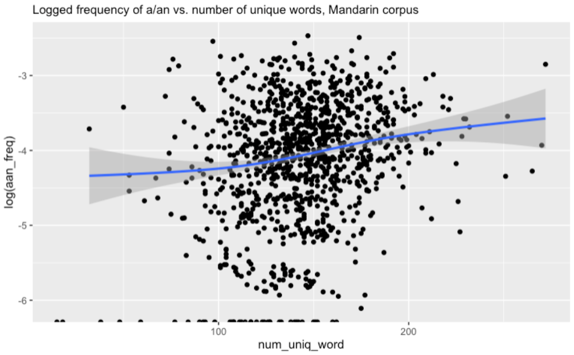


Figure 10: Logged frequency of a/an vs. number of unique words, L1 Mandarin

4.2.2 ”The” vs. ”a”/”an”

We first looked at the frequency of ”a”/”an” predicted by number of unique words. Fig. 10 shows a slight positive correlation, which implies that more advanced English speakers has a higher frequency of these indefinite articles.

We then plotted log-transformed ratio between ”the” and ”a”/”an” in Mandarin L1 corpus in Fig. 11. The average ratio between these two articles decreases as the author's score level goes up. When an English speaker becomes advanced in the language, they should use these two articles to almost the same extent and therefore be able to freely change between definite and indefinite phrases. It is therefore not surprising that authors of essays with higher scores have a lower than more even ratio between the two articles.

From the plot, we also see that the lower-scored essays have a higher ratio between ”the” and ”a/an”, and the variance is greater. Combining this and Fig. 10, the higher ratio of lower-scored essays could mean that beginner English speakers tend to use a bit more ”the” than ”a” or ”an”. This could imply that, as Mandarin speakers are early in their English learning process, they are not able to distinguish between the concepts of definiteness and indefiniteness, and they tend to begin by overusing ”the”. The higher variance of the ratio might also be due to the fluctuating usage when the user is still learning the article system. Since unique and anaphoric definites are encoded in different ways in Mandarin but with the same marker ”the” in English, it is possible that beginner English learners from Mandarin L1 background use ”the” as the single marker for any marked definite phrases before learning about the difference in function between ”the” and the indefinite markers.

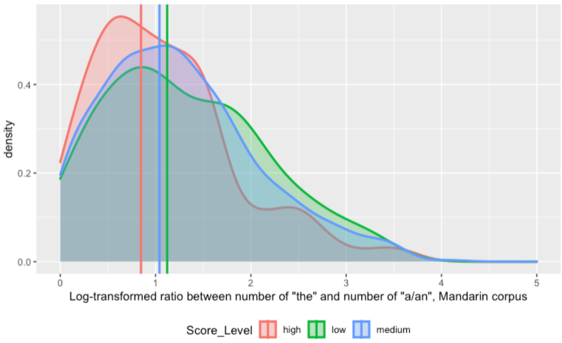


Figure 11: Logged ratio between number of ”the” and number of ”a” and ”an”, L1 Mandarin

4.2.3 ”The” vs. Other Determiners

There is a category of determiners that we selected to be a separate category during the data preprocessing step, and we called these determiners ”other determiners.” These are words that are tagged as ”determiner” by the NLTK tagger besides the definite and indefinite articles, and this category includes ”some”, ”every”, ”all”, etc. All of these determiners have lexical counterparts in Mandarin. Since these determiners exist in Mandarin, but the articles ”the” and ”a”/”an” do not, we wanted to see how the usage of these determiners with lexical counterparts is different from that of the determiners that do not exist in a speaker’s L1 language.

We plotted the log-transformed ratio between the number of ”the” and the number of these ”other determiners” for the Mandarin L1 corpus. Fig. 12 shows that the logged ratio between these two categories of words seems to converge to around the same regardless of the essay’s score level. This seems to suggest that Mandarin L1 speakers of different proficiency levels use these determiners that have lexical counterparts in Mandarin in a similar way. This could be unsurprising due to the fact that these ”other determiners” usually have distinct semantics. For example, the word ”every” has a very specific semantic function, so beginning English students who know the word ”every” might not use it in situations that don’t convey the semantic meaning of the word itself.

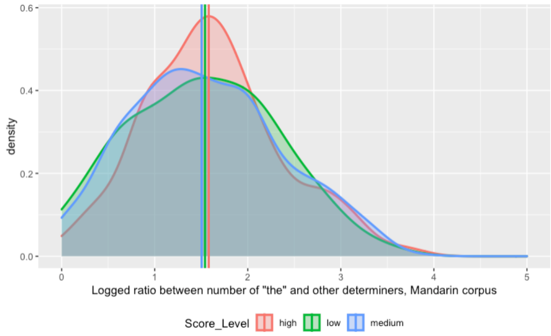


Figure 12: Logged ratio between number of ”the” and other determiners, L1 Mandarin

5. Conclusions and Future Work

In this paper, we aim to investigate the use of articles among English speakers of different levels of proficiency and from different L1 backgrounds. The data exhibits fairly high variance, so correlation between the predicted variables (eg. frequency and ratio of certain articles) and the independent variables (ie. the proxies for English proficiency) tends to be weak in general.

We propose that L2 English speakers tend to encode definiteness more frequently with their native languages’ counterparts if definite article is not present in that language. Specifically, we investigated the use of English demonstratives by L1 Mandarin speakers. It turns out demonstratives’ use does not significantly influence learner’s acquisition of the English definite article. The variances of the frequency of demonstratives and the definite article ’the’ become smaller as the essay’s score level increases, which could indicate that learners have understood the appropriate use of definiteness and demonstratives as they have become more familiar with the English and mastered it better.

We also investigated how native Mandarin speakers acquire indefiniteness along with definiteness. The acquisition of indefiniteness influences the acquisition of definiteness to a certain extent. There is a positive correlation between the frequency of indefinite articles and number of unique words in the essay (our proxy of English proficiency), and Mandarin speakers with a higher level of English proficiency have a more even ratio between the definite and indefinite articles. Beginner English learners with Mandarin L1 background seem to overuse ”the”.

We also looked at the usage of other determiners other than definite and indefinite articles in Mandarin native speakers. These determiners have their lexical counterparts in Mandarin. For L1 Mandarin speakers, they use these determiners similarly across score levels. This phenomenon suggests that they only need to familiarize themselves with these determiners other than acquiring them.

Overall, L2 learners of English learning the definite article system do acquire the concept of definiteness and indefiniteness through the course of becoming proficient in English. Since definiteness exists in Mandarin and indefiniteness does not, it is easier for L1 Mandarin learners to acquire definiteness over indefiniteness. For those determiners present in their L1, the usage does not change greatly, but it stabilizes with the person's increased familiarity and proficiency with English.

There are difficulties during the analysis process. For one, due to the nature of examining a corpus, it was not possible for us to detect instances where an article should be used in a noun phrase but is omitted by the writer – we could only see and count the instances where these articles *are* used. Investigating the omission of articles is an important and helpful step in similar projects. The lack of a control group corpus based on native English speakers' writing is also not ideal, because we did not have a baseline that we could use to contrast with our findings from the non-native speakers corpora.

We envision two main areas for potential future work on similar topics to our current project. One possible investigation could be done on spontaneous speech rather than written text by non-native speakers of English, because people usually have more time to think about their grammar while writing. The Speaking portion of the TOEFL exam could be a good candidate for such studies, and we can try to see if spontaneous speech exhibits different distributions. Another approach could be asking native speakers of English to judge the correctness of articles' use by L2 speakers of English with different levels of proficiency and varied L1 backgrounds, and use this judgment as a metric for the acquisition of the article system of English.

References

[1] Daniel Blanchard. Ets corpus of non-native written english, 2014.

[2] Roger Hawkins and Cecilia Yuet hung Chan. The partial availability of universal grammar in second language acquisition: the 'failed functional features hypothesis'. *Second Language Research*, 13, 1997.

[3] Janellan Huttenlocher, Heidi R. Waterfall, Marina Vasilyeva, and Jack L. Vevea. The varieties of speech to young children. *Developmental Psychology*, 43, 2007.

[4] Tania Ionin, Heejeong Ko, and Kenneth Wexler. Article semantics in l2 acquisition: The role of specificity. *Language Acquisition*, 12, 2004.

[5] Peter Jenks. Articulated definiteness without articles. *Linguistic Inquiry*, 49, 2018.

[6] Heejeong Ko, Tania Ionin, and Ken Wexler. The role of presuppositionality in the second language acquisition of english articles. *Linguistic Inquiry*, 41, 2010.

[7] Ewan Klein Steven Bird, Edward Loper. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

[8] Marta Tryzna. Questioning the validity of the article choice parameter and the fluctuation hypothesis: Evidence from l2 english article use by l1 polish and l2 mandarin chinese speakers. *Language Acquisition and Language Disorders*, 49, 2009.

[9] Sowmya Vajjala. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28, 2018.