

**SYNTACTIC VARIATION ACROSS PROFICIENCY LEVELS  
IN JAPANESE EFL LEARNER SPEECH**

---

A Dissertation Submitted  
to the Temple University Graduate Board

---

In Partial Fulfillment  
of the Requirements for the Degree of  
Doctor of Education

---

By  
Mariko Abe  
December 2015

Advisory Committee

Paul Nation, Advisory Chair, Victoria University of Wellington  
David Beglar, Teaching and Learning  
Martin Willis, External Examiner, Tokyo Woman's Christian University  
Yukio Tono, External Member, Tokyo University of Foreign Studies  
Brad Visgatis, External Examiner, Osaka International University

©

Copyright

2015

by

Mariko Abe

## ABSTRACT

Overall patterns of language use variation across oral proficiency levels of 1,243 Japanese EFL learners and 20 native speakers of English using the linguistic features set from Biber (1988) were investigated in this study. The approach combined learner corpora, language processing techniques, visual inspection of descriptive statistics, and multivariate statistical analysis to identify characteristics of learner language use. The largest spoken learner corpus in Japan, the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) Corpus was used for the analysis. It consists of over one million running words of L2 spoken English with oral proficiency level information. The level of the material in the corpus is approximately equal to a Test of English for International Communication (TOEIC) range of 356 to 921. It also includes data gathered from 20 native speakers who performed identical speaking tasks as the learners.

The 58 linguistic features (e.g., grammatical features) were taken from the original list of 67 linguistic features in Biber (1988) to explore the variation of learner language. The following research questions were addressed. First, what linguistic features characterize different oral proficiency levels? Second, to what degree do the language features appearing in the spoken production of high proficiency learners match those of native speakers who perform the same task? Third, is the oral production of Japanese EFL learners rich enough to display the full range of features used by Biber?

Grammatical features alone would not be enough to comprehensively distinguish oral proficiency levels, but the results of the study show that various types of

grammatical features can be used to describe differences in the levels. First, frequency change patterns (i.e., a rising, a falling, a combination of rising, falling, and a plateauing) across the oral proficiency levels were shown through linguistic features from a wide range of categories: (a) part-of-speech (*noun, pronoun it, first person pronoun, demonstrative pronoun, indefinite pronoun, possibility modal, adverb, causative adverb*), (b) stance markers (*emphatic, hedge, amplifier*), (c) reduced forms (*contraction, stranded preposition*), (d) specialized verb class (*private verb*), complementation (*infinitive*), (e) coordination (*phrasal coordination*), (f) passive (*agentless passive*), and (g) possibly tense and aspect markers (*past tense, perfect aspect*). In addition, there is a noticeable gap between native and non-native speakers of English. There are six items that native speakers of English use more frequently than the most advanced learners (*perfect aspect, place adverb, pronoun it, stranded preposition, synthetic negation, emphatic*) and five items that native speakers use less frequently (*past tense, first person pronoun, infinitive, possibility modal, analytic negation*). Other linguistic features are used with similar frequency across the levels. What is clear is that the speaking tasks and the time allowed for provided ample opportunity for most of Biber's features to be used across the levels.

The results of this study show that various linguistic features can be used to distinguish different oral proficiency levels, and to distinguish the oral language use of native and non-native speakers of English.

## ACKNOWLEDGMENTS

I would like to express my deepest gratitude to the dissertation advisory chair, Paul Nation, for his keen insights and direction throughout this long dissertation project. His enthusiasm and support helped me enormously to make this project a positive one. I am also indebted to Dr. David Beglar for guiding me all through the laborious work of dissertation writing and for deepening the discussion of the study. Dr. Yukio Tono joined my advisory committee at the stage of proposal defense and provided me with profound corpus linguistic knowledge. Dr. Akira Murakami agreeably allowed me to revise his Perl scripts for conducting this learner language analysis and also provided me with useful programming advice. Dr. Brad Visgatis spared his precious time to be a member of my final defense committee. Dr. Laurence Anthony set me on the right path at an earlier stage in the presentation of the thesis. Dr. Martin Willis supported me hugely throughout the process of completing the final manuscript by providing his thoughtful advice on various types of statistical techniques that had a major influence on this project. Finally, I wish to thank my friends at Victoria University of Wellington and Temple University Japan for their warm support and friendship.

## TABLE OF CONTENTS

	PAGE
ABSTRACT .....	iii
ACKNOWLEDGMENTS.....	v
LIST OF TABLES .....	x
LIST OF FIGURES.....	xii
CHAPTER	
1. INTRODUCTION .....	1
The Importance of Learner Corpora.....	1
Statement of the Problem .....	4
Purposes and Significance of the Study .....	6
The Audience for the Study.....	8
Delimitations .....	9
The Organization of the Study .....	10
2. REVIEW OF THE LITERATURE .....	12
Corpus-Based Learner Language Studies: Differences Between the First Languages.....	12
Corpus-Based Learner Language Studies: Differences Between Native and Non-Native Speakers of English .....	14
Corpus-Based Learner Language Studies: Differences Between English Proficiency Levels.....	19
Corpus-Based Learner Language Studies Analyzed from Multiple Aspects .....	29
Gaps in the Literature .....	37
The Purposes of this Study and Research Questions.....	39
3. METHODS .....	42

The Learners Who Contributed to the Corpus.....	42
Gender and Age.....	45
Overseas Experience .....	46
Scores on English Examinations .....	47
Instrumentation.....	48
Interviewers and Raters .....	48
English Oral Proficiency Levels.....	50
Scoring Criteria .....	52
Assessment Tasks.....	52
Corpus Data.....	54
Corpus Construction.....	54
Reference Corpus .....	56
Data Cleaning .....	57
Counting Overall Corpus Size.....	60
Linguistic Features .....	64
Part-of-Speech Tagging.....	70
Problems in the Part-of-Speech Tagging.....	74
Accuracy Rate of the Part-of-Speech Tagging .....	77
Frequency Counting .....	79
Statistical Analyses.....	80
Box-and-Whisker Plot Analysis .....	80
Correspondence Analysis .....	81
Procedures .....	84
4. RESULTS .....	87
Results of Descriptive Statistics .....	87
Interpreting the Box-and-Whisker Plots.....	101

Results of Box-and-Whisker Plots Analysis .....	105
Falling Frequency Change Pattern .....	106
Rising Frequency Change Pattern .....	108
Other Frequency Change Patterns .....	112
Results of Correspondence Analysis .....	126
Correspondence Analysis Without Native Speakers of English.....	137
Correspondence Analysis with Native Speakers of English as Supplementary Variable .....	143
5. DISCUSSION.....	148
Answers to the Research Questions .....	148
Research Question 1 .....	148
Research Question 2.....	154
Research Question 3.....	155
The Strengths and Weaknesses of Correspondence Analysis and Box-and-Whisker Plot Analysis .....	157
6. CONCLUSION.....	159
Summary of the Findings .....	159
Limitations .....	160
Suggestions for Future Research.....	163
Final Conclusions.....	164
REFERENCES CITED .....	166
APPENDICES	
A. A SUMMARY OF THE SST ASSESSMENT CRITERIA .....	180
B. PART-OF-SPEECH TAGSETS USED FOR TREETAGGER .....	183
C. TAGS OF TARGETED LINGUISTIC FEATURES.....	189
D. R SCRIPTS FOR CORRESPONDENCE ANALYSIS .....	198

E. BOX-AND-WHISKER PLOTS OF TARGETED LINGUISTIC FEATURES.....	199
--	-----

## LIST OF TABLES

Table	Page
1. Gender and Age Range of the Learners by SST Level .....	45
2. The Ages of the Learners by SST Level .....	46
3. Overseas Experience of the Learners by SST Level .....	47
4. Scores on the TOEIC and TOEFL Examinations by the Test-Takers .....	49
5. A Comparison of the Levels of the SST and ACTFL OPI (Retrieved July 11, 2011 from ALC Press Website: <a href="http://www.alc.co.jp/edusys/sst/e/index.html">http://www.alc.co.jp/edusys/sst/e/index.html</a> ) .....	51
6. Basic Discourse Tags Guideline ver. 2.1.3 (Extracted from Izumi, Uchimoto, & Isahara, 2004b) .....	55
7. Examples of Basic Discourse Tags (Extracted from Izumi, Uchimoto, & Isahara, 2004b) .....	56
8. Unnecessary Periods and Spaces Removed from the Texts .....	58
9. Descriptive Statistics for the SST Oral Proficiency Levels .....	62
10. Fifty-Eight Linguistic Features Analyzed in the Present Study .....	65
11. Grammatical Features from the Previous Studies that Overlapped with Those Analyzed in the Present Study .....	68
12. Sample POS Tagged Data .....	73
13. Learner Data Tagged with a Grammatically Incorrect POS Tag .....	74
14. Learner Data Tagged by a Grammatically Correct POS Tag .....	75
15. Examples of Problematic Cases in Assigning POS Tag .....	76
16. The POS Tagging Error Coverage of One Error in Every X Words .....	78
17. Descriptive Statistics for Level 3, 4, 5, and 6 (Normalized Frequency per 100 Words) .....	89
18. Descriptive Statistics for Level 7, 8, 9, and Native Speakers of English (Normalized Frequency per 100 Words) .....	92
19. Descriptive Statistics for Level 3, 4, 5, and 6 (Raw Frequency Data) .....	95

20. Descriptive Statistics for Level 7, 8, 9, and Native Speakers of English (Raw Frequency Data).....	98
21. Confidence Intervals Across the Levels for C7 (Second Person Pronoun).....	102
22. Confidence Intervals Across the Levels for E15 (Noun) .....	107
23. Confidence Intervals Across the Levels for J42 (Emphatic).....	109
24. Confidence Intervals Across the Levels for C9 (Pronoun <i>it</i> ) .....	110
25. Confidence Intervals Across the Levels for H23 (Infinitive) and K44 (Possibility Modal).....	114
26. Normalized Frequency Counts in One Million Words of 58 Grammatical Features for Eight Proficiency Levels .....	128
27. Raw Frequency Counts of 58 Grammatical Features for Eight Proficiency Levels.....	131
28. Coordinates for the SST Oral Proficiency Levels Ranked in the Score of Dimension 1 .....	135
29. Coordinates for the Linguistic Features Ranked in the Score of Dimension 1.....	139
30. Confidence Intervals Across the Levels for M51 (Contraction) .....	141
31. Confidence Intervals Across the Levels for L48 (Private Verb).....	142
32. Normalized Frequency Counts in One Million Words of Three Grammatical Features with Similar Frequencies of Occurrence.....	151

## LIST OF FIGURES

Figure	Page
1. Number of learners at each SST oral proficiency level ( $N = 1,281$ ) .....	52
2. Mean tokens for the SST oral proficiency levels ( $N = 1,263$ ).....	63
3. Box-and-whisker plots for C7 (second person pronoun).....	101
4. Box-and-whisker plots for J39 (downtoner).....	104
5. Box-and-whisker plots for E15 (noun).....	106
6. Box-and-whisker plots for J42 (emphatic) .....	108
7. Box-and-whisker plots for C9 (pronoun <i>it</i> ) .....	110
8. Box-and-whisker plots for H23 (infinitive) and K44 (possibility modal).....	113
9. Box-and-whisker plot F16 (agentless passive), I37 (adverb), H30 (causative adverbial subordinator), J40 (hedge), M51 (contraction), C11 (indefinite pronoun), C10 (demonstrative pronoun), L48 (private verb), and A1 (past tense). .....	116
10. Box-and-whisker plots for J41 (amplifier) .....	121
11. Box-and-whisker plots for N55 (phrasal coordination).....	123
12. Box-and-whisker plots for A3 (present tense) and A2 (perfect aspect) .....	124
13. Results of the correspondence analysis (SST oral proficiency level).....	135
14. Results of the correspondence analysis (joint column and row plot) .....	138
15. Box-and-whisker plots for M51 (contraction).....	140
16. Box-and-whisker plots for L48 (private verb).....	142
17. Results of the correspondence analysis (joint column and row plot with native speakers of English) .....	145
18. Box-and-whisker plots for C6 (first person pronoun) .....	146
E1. Box-and-whisker plots for A1 (past tense) .....	199

E2. Box-and-whisker plots for A2 (perfect aspect).....	200
E3. Box-and-whisker plots for A3 (present tense) .....	201
E4. Box-and-whisker plots for B4 (place adverbial).....	202
E5. Box-and-whisker plots for B5 (time adverbial) .....	203
E6. Box-and-whisker plots for C6 (first person pronoun).....	204
E7. Box-and-whisker plots for C7 (second person pronoun) .....	205
E8. Box-and-whisker plots for C8 (third person pronoun).....	206
E9. Box-and-whisker plots for C9 (pronoun <i>it</i> ).....	207
E10. Box-and-whisker plots for C10 (demonstrative pronoun) .....	208
E11. Box-and-whisker plots for C11 (indefinite pronoun).....	209
E12. Box-and-whisker plots for C12 (proverb <i>do</i> ).....	210
E13. Box-and-whisker plots for D13 (direct WH-question) .....	211
E14. Box-and-whisker plots for E14 (nominalization) .....	212
E15. Box-and-whisker plots for E15 (noun) .....	213
E16. Box-and-whisker plots for F16 (agentless passive) .....	214
E17. Box-and-whisker plots for F17 ( <i>by</i> passive).....	215
E18. Box-and-whisker plots for G18 ( <i>be</i> as main verb).....	216
E19. Box-and-whisker plots for G19 (existential there).....	217
E20. Box-and-whisker plots for H20 ( <i>that</i> verb complement).....	218
E21. Box-and-whisker plots for H21 ( <i>that</i> adjective complement).....	219
E22. Box-and-whisker plots for H22 (WH clause) .....	220
E23. Box-and-whisker plots for H23 (infinitive) .....	221
E24. Box-and-whisker plots for H24 (past participial postnominal clause).....	222
E25. Box-and-whisker plots for H25 ( <i>that</i> relatives in subject position).....	223

E26. Box-and-whisker plots for H26 ( <i>that</i> relatives in object position).....	224
E27. Box-and-whisker plots for H27 (WH relatives in subject position).....	225
E28. Box-and-whisker plots for H28 (WH relative in object position).....	226
E29. Box-and-whisker plots for H29 (WH relatives with fronted preposition) .....	227
E30. Box-and-whisker plots for H30 (causative adverbial subordinator) .....	228
E31. Box-and-whisker plots for H31 (concessive adverbial subordinator).....	229
E32. Box-and-whisker plots for H32 (conditional adverbial subordinator) .....	230
E33. Box-and-whisker plots for H33 (other adverbial subordinator).....	231
E34. Box-and-whisker plots for I34 (prepositional phrase) .....	232
E35. Box-and-whisker plots for I35 (attributive adjective).....	233
E36. Box-and-whisker plots for I36 (predicative adjective) .....	234
E37. Box-and-whisker plots for I37 (adverb).....	235
E38. Box-and-whisker plots for J38 (conjunct) .....	236
E39. Box-and-whisker plots for J39 (downtoner) .....	237
E40. Box-and-whisker plots for J40 (hedge).....	238
E41. Box-and-whisker plots for J41 (amplifier).....	239
E42. Box-and-whisker plots for J42 (emphatic).....	240
E43. Box-and-whisker plots for J43 (discourse particle) .....	241
E44. Box-and-whisker plots for K44 (possibility modal) .....	242
E45. Box-and-whisker plots for K45 (necessity modal) .....	243
E46. Box-and-whisker plots for K46 (predictive modal) .....	244
E47. Box-and-whisker plots for L47 (public verb) .....	245
E48. Box-and-whisker plots for L48 (private verb) .....	246
E49. Box-and-whisker plots for L49 (suasive verb) .....	247

E50. Box-and-whisker plots for L50 (seem appear).....	248
E51. Box-and-whisker plots for M51 (contraction) .....	249
E52. Box-and-whisker plots for M52 (stranded preposition).....	250
E53. Box-and-whisker plots for M53 (split infinitive).....	251
E54. Box-and-whisker plots for M54 (split auxiliary) .....	252
E55. Box-and-whisker plots for N55 (phrasal coordination) .....	253
E56. Box-and-whisker plots for N56 (independent clause coordination) .....	254
E57. Box-and-whisker plots for O57 (synthetic negation).....	255
E58. Box-and-whisker plots for O58 (analytic negation).....	256

# CHAPTER 1

## INTRODUCTION

### **The Importance of Learner Corpora**

A corpus is a computer-readable language database that is systematically compiled for linguistic analysis. Large computational databases of the written and spoken production of native speakers of English have been compiled and annotated with various types of linguistic information (e.g., part-of-speech tagging, lemma information, syntactic parsing, semantic annotation) (Gries & Berez, forthcoming), and these have enabled researchers to examine language quantitatively. One of the most well-known corpus-based grammar books, *The Longman grammar of spoken and written English* (Biber, Johansson, Leech, Conrad, & Finegan, 1999), has described the general characteristics of English grammar and also challenged the traditional view of English grammar by examining computerized linguistic data. This grammar book was based on the language use of native speakers of English. However, as corpus analyses became more and more common, the size and variety of corpora increased. Accordingly, corpus studies have been applied in the domains of learner, bilingual, multilingual, child, and teenage language use (Svartvik, 1996).

A learner corpus, like the one used in this study, is systematically collected computer-readable written or spoken natural language performance data from language learners (Biber, Conrad, & Reppen, 1998; Granger, 1994; Kennedy, 1998; Leech, 1998), and it can be considered as a database of interlanguage - an incomplete grammatical

system that language learners possess during their acquisition of the target language (Selinker, 1972). Previously, learner corpora were not typical research resources in the field of second language acquisition study probably because their construction involved the manual inputting of large amounts of data. However, the construction of learner corpora has flourished in the past two decades, particularly as personal computers became more widely used. Learner corpora can be subjected to the same corpus linguistic methods, such as tagging or parsing, as native-speaker corpora, and their construction and use have yielded numerous benefits for the investigation of learner language. By adding part-of-speech information in learner data, for example, frequency information regarding the use of infinitives (*to*-clause) can be extracted without including frequency information of the preposition (*to*). Thus, “learner corpora have a much greater potential if specific language properties have been previously identified and signaled in the corpus, that is, if the corpora have been annotated” (Díaz-Negrillo & Thompson, 2013, p. 13). In other words, corpus linguistics method has enabled researchers to investigate how learners use particular grammatical structures at particular stages of language development, including overuse and underuse (Granger, 1998b; Gilquin, Papp, & Díez-Bedmar, 2008; Granger, Gilquin, & Meunier, 2013; Götz, 2015; Ishikawa, 2013).

In addition, by annotating error information in learner data, frequency information regarding the incorrect use of linguistic features can be extracted so that researchers can investigate the areas in which learners experience difficulty with the target language. Computer-aided Error Analysis (CEA) “enables researchers and language testers to describe language proficiency on a quantitative level by way of characterizing the

frequencies, types and contexts of errors that learners commit at a certain proficiency level” (Götz, 2015, p. 2). In this way, in terms of the quantitative investigation of learner language, learner corpora are useful for examining how learners use the target language at a particular stage and what problems they encounter during the process of language learning.

Learner corpora can also be used to explore the degree to which learners deliberately avoid or do not use particular structures, what learners cannot produce accurately, as well as what they can or do produce by comparing their production with that of native speakers of the target language. Some learner corpora available to the public are accompanied by a native-speaker reference corpus. The International Corpus of Learner English (ICLE) (Granger et al., 2002; Granger et al., 2009) for example has a reference corpus, namely the Louvain Corpus of Native English Essays (LOCNESS), made up of A level essays written by British pupils (60,209 words), American university students (168,400 words), and British university students (95,695 words). The National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) corpus (Izumi, Uchimoto, & Isahara, 2004b) also has a reference corpus that consists of spoken performances produced by native speakers of English. If there is a learner corpus that includes data produced by native speakers who engaged in the same task as the learners, researchers can conduct more revealing examinations of learner language.

Additionally, researchers can compare learner corpora containing data from learners who speak different native languages, and these can be used to explore varieties

of interlanguage and reveal the influence of the first language (L1) on language acquisition. With this in mind, the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011) and the International Corpus of Crosslinguistic Interlanguage (ICCI) (Tono, Kawaguchi, & Minegishi, 2012), which both contain the Common European Framework of Reference for Languages (CEFR) level information, were released as data-sources for the cross-linguistic analysis of interlanguages.

Thus, digitized learner performance data have the potential to make contributions to studies of learner language by illuminating the tendencies of learners' language use, misuse, overuse, underuse, and zero-use in language learning discourse, and similar and dissimilar use among different L1 groups. The idea of analyzing electronic linguistic data quantitatively was most strikingly applied to the frequency-based Longman Grammar of Spoken and Written English (Biber, Johansson, Leech, Conrad, & Finegan, 1999), which graphically showed many of the defining characteristics of modern English grammar. Applying a similar approach to analyzing electronic learner language data can open up the possibility of gaining a more accurate and comprehensive understanding of interlanguage and uncovering factors influencing foreign language acquisition.

### **Statement of the Problem**

Although learner corpora have the potential to increase our understanding of interlanguage, few attempts have been made to provide systematic data about spoken language use that can be applied to language teaching and assessment materials (Pendar & Chapelle, 2008). Despite some early work in the field of SLA, relatively few

researchers have continued to be concerned with the development of oral proficiency and few have described non-native speaker interlanguage development using a range of linguistic features. The number of targeted linguistic features in previous SLA studies is limited (Biber, Conrad, & Reppen, 1998), and consequently, it is still not clear which linguistic features are useful in describing different oral proficiency levels, and this has resulted in a gap in our understanding of the characteristics of foreign language learner interlanguages.

This neglect is partly because interlanguage performance data with proficiency level information that is based on an objective rubric have not been readily available to researchers. Learner corpora are relatively recent, with the construction of such corpora increasing in the 1990's, including the well-known ICLE, which was created as a part of the International Corpus of English (ICE) made up of national and regional varieties of English that were used for comparative studies of English. However, the ICLE only contains data from high intermediate to advanced learners of English. Several corpora based on the language production of Japanese learners of English have been developed. The Japanese EFL Learner Corpus (JEFLL) (Tono, 2007) consists of essays written by junior and senior high school students. The Nagoya Interlanguage Corpus of English (NICE) (Sugiura, 2008) consists of essays written by university students. Yet, these learner corpora cannot be used to track how learner language changes from lower to higher proficiency levels because they do not contain proficiency level information.

However, there is more to corpus analysis than corpora. The results of corpus analysis depend not only on the size and quality of the corpora used, but also on the tools

and procedures used to extract data from the corpora. The deficiencies in research methods used on learner corpora have been pointed out by previous researchers. Corpus linguistics techniques or language processing techniques, such as the automatic detection of relevant linguistic features, have not been used to their full potential, and thus has affected the study of the nature of learner language (Asención-Delaney, 2015; Alexopoulou, Geertzen, Korhonen, & Meurers, 2015; Biber, Conrad, & Reppen, 1998; Granger, Kraif, Ponton, Antoniadis, & Zampa, 2007; Gries, 2009, 2010; Hiayang & Lu, 2013; Lu, 2010; Meurers, 2009, 2013, forthcoming; Myles, 2005; Pendar & Chapelle, 2008). For that reason, I need to partially remedy these deficiencies.

### **Purposes and Significance of the Study**

The first purpose of this study is to identify linguistic features that can be used to describe variation across the different oral proficiency levels of a wide range of Japanese EFL learners. Linked with this purpose, the second purpose is to address the lack of SLA studies profiling interlanguage levels using multiple linguistic features. Thus, the present study concerns the field of foreign language assessment. A general profile of oral performance can be used as a benchmark by language learners to evaluate their own speaking performance. Test developers can use it to evaluate L2 oral proficiency test results, to develop and refine rating scales and speaking assessment tasks, and to train examiners and raters. The Cambridge English Profile Programme has begun the job of identifying so-called criterial features for the proficiency levels specified by the Common European Framework of Reference (CEFR) with the assistance of the Cambridge Learner

Corpus (CLC) (Hawkins & Buttery, 2009, 2010). If the distinguishing features, for example, those specific to certain levels of Japanese EFL learners are identified, they can be similarly applied to creating assessments that are more appropriate for scaling learners' performance (Tono, Kaneta, & Doi, 2010).

The third purpose of this study is to provide methodological solutions to the problems that have been pointed out by previous researchers. These include what kinds of analyses and linguistic feature sets to use, and what kinds of data are needed for such analyses. For example, Alexopoulou, Geertzen, Korhonen, and Meurers (2015) stated that "Manual data processing needs to be complemented by automated data processing if the full size of a resource of half a million scripts is to be exploited. Automated Natural Language Processing (NLP) of big learner data is, therefore, vital (Granger et al. 2007; Meurers 2009)" (p. 99). This type of approach is relatively new in the field of second language acquisition because it was not feasible until digitized learner performance data containing oral proficiency level information became available. Learner data with oral proficiency levels now allows researchers to use language-processing techniques to obtain large amounts of frequency information on language use by learners at various levels, and this allows researchers to examine the characteristics of learner language variation across proficiency levels and to provide the potential for producing an overview of interlanguage use.

## **The Audience for the Study**

The main audience for the present study is researchers investigating spoken language acquisition because the results can indicate linguistic features that learners at particular oral proficiency levels have acquired and might be ready to acquire. Other results include a more comprehensive understanding of interlanguage variation across oral proficiency levels and the differences in language use between native and non-native speakers of English. Therefore, the results can lead researchers to prioritize linguistic features that should be investigated in future studies. I hope to contribute to our understanding of learner language. However, a general description of the linguistic forms used by learners at a particular oral proficiency level cannot be a guide for sequencing linguistic items in a syllabus or teaching materials until future research has shown that such an approach is feasible.

The second audience for this study is language assessment professionals and particularly individuals involved in the assessment of oral English proficiency. They can benefit from the present study because a learner corpus study, which is based on data gathered in an oral interview, can clarify the grammatical, lexical, and discourse forms used by individuals at different oral proficiency levels. While it is still unclear whether the findings can be applied to high-stakes speaking assessments, they have the potential to be used to validate speaking tests, to develop rating scales and speaking assessment tasks, and to train examiners and raters.

## **Delimitations**

There are four delimitations in using a corpus-based approach. The first delimitation concerns the data collection method. Experimental methods, such as think-aloud protocols, grammatical judgment tests, and retrospective learner reports, have been widely used to examine the use of particular linguistic features in SLA studies (Ellis, 2008). These data collection methods target particular linguistic features, but a corpus-based research method uses data gathered during an interview or from writing a composition (Granger, 1998a; McEnery, Xiao, & Tono, 2006). Therefore, the data used in this study are likely to have been influenced by the interviewers or the elicitation tasks. This is both a weakness and strength. As a weakness, I cannot be sure what role the task and interviewers played in the production of certain features. For example, a particular elicitation task that is used to elicit oral production can emphasize some morpho-syntactic structures and de-emphasize others. As a strength, the use of standard task rubrics and standard interview procedures provides a common framework for the language produced by the different learners, at least partially controlling an influential performance variable—the nature of the task.

The second delimitation is that because of the collection method, the data are generally not made up of multiple performances gathered from the same individuals over an extended period of time. Test-takers at various oral proficiency levels are included in this study, but researchers cannot capture the process of language development for each individual because the data were gathered at one point in time rather than longitudinally. Thus, the learner corpus in the present study cannot be used for conducting longitudinal

studies. However, because the data of every learner is classified using their oral proficiency level, it is possible to see snapshots of performance by different learners on similar tasks across a wide range of levels.

The third delimitation concerns the retention rate of automatically retrieved complex grammatical features. Syntactic information is essential in calculating the frequencies of complex structures and in distinguishing frequency counts of words that can be categorized as more than one part-of-speech. Thus, it is necessary to prepare a well-annotated corpus to identify grammatical features and to carry out lexical searches to remove erroneously extracted items manually while maintaining a reasonable retention rate of accurately tagged features (Gries, 2008).

The last delimitation concerns the pedagogical implications of this study. The results of the present study cannot be directly applied to classroom use because the results do not indicate which linguistic items should be taught at earlier or later stages of language learning or how easy or difficult those items are to acquire. The study is designed to understand differences in learner language across proficiency levels and between native and non-native speakers of English. It is not possible to verify why certain characteristics of learner language occur at certain stages as was claimed in Tono (2000).

### **The Organization of the Study**

In Chapter 2, Review of the Literature, I review previous work in the learner corpus literature and identify gaps in previous studies. At the end of the chapter, I present

the research purposes and research questions that guide this study. In Chapter 3, Methods, I describe the learners, instrumentation, corpus data, statistical analyses, and research procedures used in the study. In Chapter 4, Results, I present the linguistic features that can be used to characterize oral proficiency groups of Japanese EFL learners. In Chapter 5, Discussion, I provide interpretations of the results. In Chapter 6, Conclusion, I briefly summarize the findings, discuss the limitations of the study, provide suggestions for future research, and make concluding comments.

## **CHAPTER 2**

### **REVIEW OF THE LITERATURE**

In this chapter I first review the corpus-based learner language literature from the viewpoint of differences between L1s, differences between native speakers of English (NSE) and non-native speakers of English (NNSE), and differences between English oral proficiency levels. I also review the studies in which multiple linguistic features were used to understand the overall patterns of language variation. I then point out the analytical gaps in previous studies and describe the contribution of the present study. Finally, the purposes of the study and research questions are provided at the end of this chapter.

#### **Corpus-Based Learner Language Studies: Differences Between the First Languages**

As the construction of learner corpora flourishes all over the world, researchers have begun to investigate varieties of learner languages in more detail (Ellis, 2008). This approach to research is called Contrastive Interlanguage Analysis (CIA), a research paradigm shift instigated by Granger (1996). She divided this approach into two types (Granger, 1998b, 1988c). One is to examine the first language (L1) based factors that influence the development of learner language. It involves comparing learner corpora which consist of data from different L1 learners (e.g., French, Japanese, and Chinese). For example, in order to conduct a CIA study regarding written performance, researchers can use the International Corpus of Learner English (ICLE), one of the most recognized

corpora of learner English in the world. This 3.7 million-word corpus consists of 500- to 1,000-word argumentative essays written by high intermediate to advanced learners of English from 16 native language backgrounds, including Japanese. Other than ICLE, there are large commercial written learner corpora, such as the Cambridge Learner Corpus (CLC) and Longman Learner Corpus (LLC), but they are largely restricted to the internal use of the publishers in their publications and in educational research. For spoken language, researchers can use the Louvain International Database of Spoken English Interlanguage (LINDSEI) to conduct a CIA study. This spoken corpus was launched in 1995 by the same institution that constructed the ICLE. It consists of 554 interview transcripts totaling over one million words of language from learners from 11 different native language backgrounds.

Many previous studies have used the ICLE. Ringbom (1998) investigated whether there were many differences in vocabulary use frequencies among English learners with different L1s, and found that different L1 learners overuse the top 30 to 100 most frequently used words. Altenberg (2002) argued that Swedish English learners' overuse of the causative *make* was caused by L1 transfer, especially because of cross-linguistic similarity, comparing it with French learners' underuse of the form. Aijmer (2002) found that the overuse of modal auxiliaries, modal adverbials, and lexical verbs with modal meaning was a shared characteristic of French, German, and Swedish L2 writers, and partly reflected developmental and interlingual characteristics of learner language use. Nesselhauf (2005) explored the verb-object-noun collocations (e.g., *take a break*) of advanced German English learners, and showed that the L1 has a greater influence on the

production of collocations than has been shown in previous smaller scale non-corpus-based studies. Biber and Reppen (1998) investigated the use of complement clauses, and reported that four learner groups, French, Spanish, Chinese, and Japanese, shared particular patterns in the use of complement clauses; (a) *that*- and *to*-clauses were significantly common, (b) *-ing* and WH-clauses were uncommon, and (c) the use of the *that*-clauses in learners' essays was similar to their use in the conversations of English speakers.

By comparing digitized learner performance data among different L1 backgrounds, researchers can gain vast amounts of frequency-based information on vocabulary or sentence structures, and this information enables researchers to reveal language use patterns of different L1 learners, and thus gauge the effect of their first language on L2 use (Ortega, 2009). Researchers can also shed light on whether certain features of learner language are universal phenomena or unique developmental characteristics peculiar to a specific first language (Gilquin, Papp, & Díez-Bedmar, 2008).

### **Corpus-Based Learner Language Studies: Differences Between Native and Non-Native Speakers of English**

As was mentioned in previous section, one of the purposes of CIA is to examine the L1 based factors that influence the development of learner language. Another purpose of CIA is to examine language use differences between native and non-native speakers of English, so that researchers can reveal the overuse, underuse, and misuse tendencies of

language learners (Gilquin, Papp, & Díez-Bedmar, 2008). The NICT JLE corpus, for example, has a reference corpus consisting of production data from tasks also completed by native speakers of English, and this reference corpus enables researchers to compare the spoken production of native and non-native speakers performing the same tasks.

However, this type of comparison between native and non-native speakers of English might distract researchers from seeing the uniqueness of learner language systematicity. Comparing second or foreign language learners with L1 learners has been seen to result in a misperception of how second language learners develop their interlanguage. It is argued that learners' interlanguage needs to be viewed in terms of its own system not using a native speaker system. Applying the L1 system to second language analysis is given the derogatory title of "the comparative fallacy" (Bley-Vroman, 1983; Lakshmanan & Selinker, 2001). The interlanguage systems of EFL learners include what could be called ill-formed language use patterns that deviate from the target language norms, but these systems should be considered worthwhile objects of study in order to understand how learner language develops.

To a small degree the comparative fallacy is adopted by the present study. However, the frequency information on the language use of native speakers is not used to investigate how learner language is deviant from the target norm in this study. It is used to understand the characteristics of learner language and native speakers of English. Frequency analysis aims to reveal the internal systematicity of language learners without involving judgments of right or wrong and thus largely avoids the comparative fallacy (Ellis & Barkhuizen, 2005). During the process of language learning, it is useful for

learners to understand the norms of the target language and the characteristic differences between the interlanguage and the target. It is also necessary for researchers to understand the differences in order to develop appropriate language assessment materials. Native-speaker/non-native-speaker comparisons can provide useful insights into language learning and teaching.

Following the approach of comparing corpora of native and non-native speakers of English, Lorenz (1998) found that German learners of English overused intensified adjectives (e.g., *important, good, successful*), and it was suggested that this frequent use of intensified adjectives indicates the linguistic immaturity and non-native-likeness of the learners. The number of attributive adjectives correlates with linguistic maturity in the writing of native speakers of English, but it goes in the contrary direction with foreign learners of English. Interestingly, both native and non-native learners of English chose to use more predicative adjectives than attributive intensification, but still the older learners used more attributives than the younger learners. Furthermore, the learners attempted to pack too much information into their sentences by using adjective intensification in thematic places (e.g., *A highly specified and specialized training for journalists is for that reason one demand. I thought that my absolutely authentic Rock music should hit the charts in seconds.*).

Native-speaker and non-native-speaker comparisons have also been directed towards multiword units. De Cock et al. (1998) compared how native and non-native speakers of English used prefabricated expressions, and noted that advanced learners used them differently from native speakers of English in terms of frequency. Native

speakers of English use vagueness markers (e.g., *sort of, kind of*) with verb phrases, but non-native speakers of English combine them with noun phrases. Non-native speakers of English can use these markers by using words borrowed from their L1, for example French, to fulfill pragmatic functions (e.g., *sort of braderie, sort of vapeur*). Granger (1998c) examined the use of collocations and formulae and found that learners underused amplifiers (e.g., *perfectly natural, closely linked with*) and that L1 transfer can influence the use of collocations. Similarly, Howarth (1998) investigated verb-object collocations and concluded that non-native speakers underused restricted collocations and idioms (e.g., *give the impression, curry favour*). Laufer and Waldman (2011) examined verb-noun collocations and found that learners produced far fewer collocations than native speakers of English. They also found that the use of collocations increased as proficiency level rose but interlingual errors remained even at advanced levels.

Other than these studies regarding lexical aspects, there are studies focusing on the frequencies of grammatical patterns. Aarts and Granger (1998) compared the sequential use of part-of-speech by language learners with that of native speakers of English, and revealed similar overuse and underuse tendencies among Dutch, Finnish, and French learners. Interestingly, learners with different L1s shared the same patterns, but they deviated from the native speakers. Prepositions were included in the four most frequently used trigrams (e.g., preposition + article + noun, article + noun + preposition) of the native speakers, but foreign language learners underused them. Learners also underused nouns, conjunctions followed by nouns, and prepositions followed by *-ing* verbs, but they overused connectives, adverbs, auxiliaries, and pronouns. The researchers

also succeeded in distinguishing global and local interlanguage patterns, indicating that advanced interlanguage contains more L1 specific features than universal ones. For example, there were three types of structures which were overused by French learners of English: (a) infinitive used as an adverbial of purpose at the beginning of the sentence (e.g., *To answer* the question, ...), (b) infinitive used as a subject instead of a gerund at the beginning of the sentence (e.g., *To live* in the same nation ...), and (c) coordinated marked infinitives (e.g., a real opportunity to develop and *to find* new outlet). In addition to this study, Granger and Rayson (1998) conducted word category automatic profiling and found that advanced French learners of English overused (a) indefinite articles, (b) most indefinite determiners, (c) most indefinite pronouns, (d) coordinating conjunctions *but* and *or*, (e) some complex subordinators, (f) short adverbs of native origin used for place and time, (g) auxiliaries, and (h) infinitives. However, on the other hand, the learners underused (a) definite articles, (b) the coordinating conjunction *and*, (c) most subordinators, (d) most prepositions, (e) most adverbial particles, (f) *-ly* adverbs, (g) nouns, and (h) *-ing* and *-ed* participles.

The Contrastive Interlanguage Analysis (CIA) approach has been carried out to investigate a wide range of lexical and grammatical topics, such as modals (Aijmer, 2002), collocations (De Cock et al., 1998; Granger, 1998c; Howarth, 1998; Laufer & Waldman, 2011), adjective intensifiers (Lorenz, 1998), sequential use of part-of-speech (Aarts & Granger, 1998), and word category (Granger & Rayson, 1998). By comparing the output of native and non-native speakers of English, researchers come to illuminate differences in language use that reveal the distinguishing characteristics of native- and

non-native likeness, tendencies toward over- and underuse of particular forms. The studies reviewed above show clear evidence of these tendencies, with the particularly interesting finding that although the L1 plays an important role in degree of use of a feature, there are also instances of underuse and overuse that are common to learners from a variety of language backgrounds. As Ellis (2008) noted, corpus-based Comparative Interlanguage Analysis has set a new research standard in second language acquisition studies.

### **Corpus-Based Learner Language Studies: Differences Between English Proficiency Levels**

The CIA studies comparing corpora of different L1 learners or native and non-native speakers of English showed tendencies of learner language use regarding particular linguistic features. However, there are few in-depth examinations of language learners' performance across different proficiency groups. If researchers have a learner corpus with information regarding English proficiency groups, they can assess hypotheses about language acquisition theory. Tono (2006), for example, verified the morpheme acquisition order study of Dulay and Burt (1973) using the written production of Japanese learners of English. He found that Japanese learners have difficulty acquiring articles, but not in acquiring the possessive marker *s* at an early stage. Similarly, Izumi and Isahara (2004) used the spoken production of Japanese learners of English to verify the morpheme acquisition order of Dulay and Burt. They compared their results with that

of L1 Spanish speakers and suggested that Japanese learners acquire articles and plural -s at later stages.

Other than these verification studies, learner language in different proficiency levels can be used for investigating how learner language changes across the levels. In general, it is difficult to collect data by following the linguistic development of a group of learners over time, but researchers can analyze the data of learners at the same oral proficiency levels. As pointed out in the delimitations section, this type of data does not constitute a genuine longitudinal study, but is useful in observing how interlanguage changes (Granger, 1998a). Using such cross-sectional data, some studies investigated how learner language changes across oral proficiency levels. Hasselgren (2002), for example, examined the speech of 14- and 15-year-old Norwegian learners of English taking a spoken interaction test, as well as that of a British control group. She found that learners at higher proficiency levels used a greater quantity and variety of words and phrases (e.g., *well, you know, sort of*) that facilitate the flow of conversation compared with learners at lower proficiency levels, and she concluded that these words and phrases can be indicators of oral fluency for Norwegian learners of English.

Housen (2002) investigated the development of the basic morphological categories of the English verbal system using a 230,000-word spoken learner corpus. His data consisted of 46 young Dutch and French L2 learners of English categorized into four oral proficiency levels. In order to examine the acquisition of the verbal system (e.g., base form, simple present form, present participle form, regular and irregular past participle), learners were asked to talk about past experience and future plans, describe

pictures, and retell films and picture stories through informal free conversation and a semi-guided speech task. He found that there was significant variation among individual learners in using verbal forms. He also found that L2 learners mainly underused -en and -ed markers compared to native speakers of English. Additionally, he showed that L2 learners come to use variety of forms before they use of English verbs well. There were significant individual differences among the learners from the same oral proficiency level and between learners with different L1s. As a result, he concluded that larger corpora with longitudinal data from individual learners were necessary for investigating overall frequencies and patterns of language use distribution.

In contrast to the studies that used cross-sectional data, Crossley, Salsbury, and McNamara (2010a, 2010b, 2011) used the same year-long longitudinal L2 spoken corpus which consisted of data from six learners of English. Their learner corpus data did not include oral proficiency levels, but all six learners belonged to the lowest proficiency level of a six-level test and their TOEFL scores increased during the experiment. They stated in these three studies that different types of lexical indices can be used to predict the oral proficiency levels of language learners. Crossley et al. (2010a) found that at the first stage L2 learners use less frequent concrete words (e.g., words that indicate a particular person or an object) and then begin to use more frequent polysemous words (e.g., *know*, *think*) as the time of learning English increases. They suggested that ambiguities contained in the various senses of polysemous words cause this order, and they concluded that the use of polysemous words can be indicators of L2 oral proficiency. Similarly, Crossley et al. (2010b) examined the semantic relations between words. They

found that learners come to produce more accurate semantic connections and more semantically similar terms (e.g., *feel, worry, think*) as they make progress in acquiring English. As a result of this study, they stated that closer semantic similarities between the speech segments develop as the oral proficiency level rises.

Crossley et al. (2011) examined the use of lexical bundles. They found that L2 learners used common lexical bundles more frequently than uncommon ones, and these common lexical bundles were based on pronouns, especially on first person pronouns (e.g., *I am, I want*), which shows a tendency for speaker-centered L2 speech. They claimed that the increase of L2 oral proficiency can be measured by the frequency increase of lexical bundles.

The frequency of bundles of parts-of-speech can also be used as markers of proficiency levels. Tono (2000) used the written production of Japanese learners of English to examine the relationship between school year and part-of-speech (POS) tag sequences. He found that some trigrams (e.g., verb + article + noun, preposition + article + noun, article + noun + punctuation) were used with fairly high frequency across the school-year group. He also found that preposition-related trigrams were underused by Japanese learners of English, which supports the results of Aarts and Granger (1998) that examined the language use of Dutch, Finnish, and French learners of English. Tono stated that Japanese learners of English are able to produce a prepositional phrase (preposition + article + noun), but they have a difficulty in connecting a prepositional phrase with a noun phrase or verb phrase which begins with a verb. Unlike the low frequencies of preposition-related trigrams, verb-related trigrams were highly used. Thus,

he concluded that this tendency is caused by the low frequency of preposition- and noun-related trigrams. In order to explore the relationships between part-of-speech tag sequences (i.e., verb-, preposition-, noun-related trigrams) and different school-year groups he conducted correspondence analysis. He found that learners at the beginning level were likely to use verb-related trigrams with high frequency, learners at lower-intermediate and advanced levels used noun-related trigrams with high frequency, and learners at the highest level used preposition-related trigrams with high frequency. He also found that article- and modal auxiliary-related trigrams were fairly less frequently used, and he suggested that these two POS tag sequences can be used as discriminatory features to distinguish lower and higher school-year groups. The writing of Japanese EFL learners develops in the order of noun-, verb-, and preposition-related POS tag sequences, so that complex prepositional phrases can be used as “one of the most salient characteristics of fully developed interlanguage” (p. 335).

In contrast, Kobayashi (2007b) investigated the oral performance of Japanese EFL learners, by analyzing word-class distribution using correspondence analysis. This is one of several studies that used the same learner corpus, the NICT JLE corpus, and the same statistical method as in the present study. The data of every test-taker was classified by oral proficiency level, so that it was possible to analyze how specific linguistic features were used at particular oral proficiency levels. In spite of using different production mode data, he came to a similar conclusion to Tono (2000); lower oral proficiency learners made greater use of nouns (e.g., proper noun, numeral noun) and then even greater use of verbs (e.g., *having*, *was*). It seems that the spoken production of

Japanese learners of English develops in the order of nouns to verbs, but interestingly, in the case of L2 learners of Spanish, learners began producing more verbs than nouns, and more nouns than adjectives (Marsden & David, 2008). The similar findings for oral and written use for Japanese learners are reassuring and suggest that there might be little differentiation in their spoken and written styles.

In a series of study using the NICT JLE corpus, Kobayashi (2007a) investigated whether the 100, 75, 50, and 25 most frequently used words in the learner corpus could discriminate learners at different oral proficiency levels. He combined correspondence analysis and cluster analysis to specify indices of learner language development. The top 25 frequently used words, particularly function words (e.g., articles, prepositions, conjunctions), clearly distinguish the oral proficiency levels of the NICT JLE corpus.

In another study, Kobayashi (2010) used the same statistical analysis, correspondence analysis to examine the NICT JLE corpus, and found that some linguistic features that can be closely associated with lower proficiency learners; for example, (a) vocabulary that is closely related to the topic of the interview, (b) expressions related to requests and the wishes of the speakers, and (c) present tense verbs. He also concluded that some other linguistic features are used more frequently by higher proficiency learners; such as (a) adverbs that express the assessment and attitude of the speakers, (b) past tense verbs, and (c) function words (e.g., *then, if, that, as*) that extend the sentence structure.

The work with Japanese learners using the NICT JLE corpus shows that the investigation of language use at various proficiency levels is a fruitful one, and the

consistency of the results of several studies suggest that the corpus is large enough to obtain reliable results.

Other than these learner corpus studies which aimed to investigate digitized learner performance data at different proficiency levels, some studies have used a different approach to show that grammatical aspect or syntactic complexity can be indices to predict the proficiency levels of language learners. Meunier and Littre (2013) investigated the potential of combining learner corpus research with experimental studies and found that time predicted decreases in tense and aspect errors and that the progressive aspect was problematic for advanced French learners of English. As another example of exploring the potential of using learner corpus study to profile the development of learner language, Vyatkina (2013) focused on individual developmental pathways to examine syntactic complexity in the writing of L2 German learners. It was found that there is a general developmental trend in the increase of the frequency and range of syntactic complexity features (e.g., coordinated, nominal, and nonfinite verb structures). The researcher also tracked two learners at the introductory level and succeeded in pinpointing the time when the target features emerged.

Regarding syntax complexity, Haiyang and Lu (2013) compared the language use of Chinese EFL learners of English and native speakers of English using ten syntactic complexity measures (e.g., unit length, amount of subordination and coordination, degree of phrasal sophistication). Most of these measures regarding subordination showed significant differences between the learners and native speakers of English and between different learners' proficiency levels (i.e., lower and higher two proficiency groups which

are based on university school year). However, the degree of coordination did not show differences. Only the number of coordinate phrases per T-unit showed significant differences between the native and non-native speakers of English. In addition to their syntactic complexity study, Norris and Ortega (2009) suggested the possibility that coordination can be an index for novice learners, subordination for intermediate learners, and complexity via phrasal elaboration (e.g., grammatical metaphor) for advanced learners.

In addition to the studies that investigated what learners can produce, there are some studies that focused on erroneous use of learner language and have found that learner errors can be used to predict the oral proficiency levels of language learners. Error Analysis (EA) studies were at the center of Second Language Acquisition (SLA) studies in the 1970's, but changed their focus as other approaches, for example, Computer-aided Error Analysis (CEA), advocated by Dagneaux, Denness, and Granger (1998), were developed to analyze learner errors. While both EA and CEA are data-oriented approaches, CEA entails storing and processing enormous amounts of data, enabling researchers to examine the general characteristics of language learning and the difficulties that learners face. Consequently, some researchers investigated how learner language changes and develops in terms of learner errors using learner corpora (Abe, 2007a, 2007b, 2007c; Dagneaux, Denness, & Granger, 1998; Götz, 2015; Granger, 1999; Kaneko, 2004; Thewissen, 2013).

Abe (2007a) focused on the association of errors with different oral proficiency levels to investigate how error types can characterize the use of language. She used a

learner corpus, the NICT JLE corpus, and manually annotated errors with 31 different types of error information. She found that verb-related errors (e.g., agreement, aspect) were more likely to be made by novice learners and noun-related errors (e.g., nominal case, nominal vocabulary) by advanced learners. The accuracy rate for articles increased considerably, from 58% to 75% across oral proficiency levels as proficiency developed, but nonetheless, the article clearly remained a problematic item for Japanese learners compared with other linguistic features. On the other hand, some linguistic errors have the potential to disappear as language learning progresses, for example, (a) errors involving prepositions associated with verbs (e.g., Tom's teacher accused him \*about cheating.), (b) verbal agreement errors (e.g., there \*are a cat), (c) verbal aspect errors (e.g., The people \*weren't knowing the reality.), and (d) nominal inflection errors (e.g., \*childerens), and they can be used to predict the oral proficiency levels of language learners.

Additionally, Abe (2007c) described the errors of Japanese EFL learners' written production in terms of (a) part-of-speech, (b) three error types (misformation, missing, unnecessary), and (c) English proficiency level. Misformation errors are errors that use the wrong part-of-speech or use the wrong word form (e.g., I am very *interesting* in your company.), missing errors are errors that fail to add necessary words (e.g., There is book on the desk.), and unnecessary errors are errors that add unnecessary words (e.g., We will visit *to* you again.). She described the tendencies for each error type to be related to part-of-speech and how its frequency has changed as follows: (a) misformation errors were related to the parts-of-speech with a high accuracy rate (e.g., pronouns, adjectives,

adverbs, nouns, and verbs), and their frequency decreased as proficiency developed, (b) missing errors had a strong relationship with the parts-of-speech that learners gradually came to use correctly (e.g., prepositional complements and articles), and their error frequency change patterns were complicated, and (c) unnecessary errors were largely related to conjunctions, and there were few differences in frequency change patterns through the learning stages, but the frequency of such errors gradually decreased as the learners' proficiency increased.

These detailed examinations of error categories suggested that some errors have common developmental patterns, while others vary considerably across proficiency levels and do not gradually disappear. Thus, the findings supported the assumption that errors provide information regarding the learners' current state of interlanguage development, as Corder (1967) argued, and that errors can distinguish learners' linguistic competence. However, on the other hand, Abe (2007a) also found that in some cases, the development of learner language is likely to be much better understood when it is focused on correct language use as well as on erroneous language use (Ellis, 2008). There are some categories of errors (e.g., *voice, modal verb, finite and nonfinite verb, verbal form, verb complement*) that cannot be simply explained by the patterns of what learners do wrongly but can be more clearly explained by what learners do correctly.

By using written and spoken learner corpora, researchers have examined a wide range of linguistic features to describe the development of learner language. They have identified various types of linguistic items that can discriminate learner language between different proficiency groups, such as phrases that facilitate the flow of conversation (e.g.,

*well, you know, sort of*), (Hasselgren, 2001), lexical indices (Crossley, Salsbury, & McNamara, 2010a, 2010b, 2011), part-of-speech sequence (Tono, 2000), part-of-speech (Kobayashi, 2007b; Marsden & David, 2008), frequently used words (Kobayashi, 2007a), present tense verbs, adverbs that express the assessment and attitude of the speakers, and function words (Kobayashi, 2010), and syntactic complexity (Vyatkina, 2013; Haiyang & Lu, 2013; Norris & Ortega, 2009). Additionally, erroneous uses of linguistic items can also be used to distinguish proficiency groups, such as noun- and verb-related errors (Abe, 2007a). However, these cross-sectional and longitudinal learner language studies differ widely in terms of targeted linguistic features, and this has led to the awareness that to describe learner language variation a more unified and comprehensive list of linguistic features is needed.

### **Corpus-Based Learner Language Studies Analyzed from Multiple Aspects**

As shown in the previous section, learner corpus studies have usefully contributed to the understanding of various aspects of interlanguage. Previous studies attempted to confirm the use of linguistic features characterizing different proficiency levels. Nevertheless, there were few attempts to describe the same learner language production using an extensive list of features. Biber (1988) however is a study that used a computational linguistic database to explore language variation from using multiple features. He used a specially written language analysis program to extract large amounts of language use frequency information across a variety of corpora, and then analyzed that frequency information using factor analysis to investigate variation between varieties of

spoken and written language of native speakers of English. This classic study drawing on the processing power of computers used multi-dimensional analysis, and this analysis was applied to a wide range of other studies, from cross-linguistic variation studies (e.g., Biber, 1995) to learner language variation studies.

However, some aspects of his multi-dimensional analysis approach came under criticism. For example, his choice of linguistic features was criticized for not including discourse features such as adverbials used to signal the organization of text (Ghadessy, 2003). The quality of Biber's analysis primarily depends on the features he has chosen (and not chosen) to analyze. Additionally, it needs to be borne in mind that Biber needed a list of features that a computer could search for, and this placed some restrictions on what could be chosen (Altenberg, 1989). For instance, in a part-of-speech tagged corpus, a computer can easily search for linguistic features such as nouns, passive voice, and relative clauses, while it cannot search for the zero use of some features, erroneously used features, or features such as phrasal verbs which require human analysis to identify. The list was also criticized because he used a mixture of different types of linguistic features, covering the syntactic, semantic and lexical levels including word length, and type-token ratio (TTR) (Nakamura, 1995). This criticism however ignores Biber's purpose of trying to find a wide range of potentially discriminating features. In addition to these points, his linguistic features' list was criticized in terms of statistical analysis. This was because the choice and the number of variables (i.e., linguistic features) can affect the outcome of the factor analysis, the statistical analysis which was used to conduct multi-dimensional analysis (Altenberg, 1989; Lee, 2003; McEnery & Hardie, 2012). Further to criticisms

regarding linguistic features, the statistical analysis used in his study (i.e., factor analysis) was criticized. The wide variety of procedures with various statistical parameters, such as rotations and number of factors, can have an influence on the number and characteristics of the resulting dimensions (Nakamura, 1995; Lee, 2003; McEnery & Hardie, 2012). In addition, the use of binary opposition (e.g., formal or informal, concrete or abstract) was criticized by McEnery and Hardie because “there is also the possibility of gradience—rather than a binary opposition” (p. 105). Ideally, multidimensional analysis should consider gradual change across proficiency levels of language learners as well as binary oppositions.

In spite of the criticisms of this method of describing register variation using multiple linguistic features, it was applied to other studies, for example, Biber and Conrad (2009). The salient differences in English text were described in their register variation study, including spoken interpersonal registers (e.g., conversation, university office hours, service counters), general written registers (e.g., newspaper writing, academic prose, fiction), and electronic registers (e.g., e-mail, internet forums, text messages) to show how multi-dimensional analysis can be used to show the clustering of linguistic features in a wide range of uses of English. This study included some linguistic features which were not included in Biber (1988), such as prepositional phrases and complement clauses controlled by verbs and nouns, the zero use of certain grammatical features, such as articles, complementizer *that*, and relative pronouns. However, these additional features were not included in the present study, largely because of the way that the corpus used was tagged.

Many other studies have used the same kind of analysis and his list to successfully examine linguistic variation. Connor-Linton and Shohamy (2001) suggested that Biber's multi-dimensional approach can be used to examine the lexical and syntactic variation of language learners and to compare the communicative competencies of speakers in different EFL oral proficiency groups. de Mönnink, Brom, and Oostdijk (2003) showed that Biber's multi-dimensional approach can be applied to a different set of linguistic data, in their case the fully parsed ICE-GB corpus. Murakami (2009) used it to explore language variation in Asian ELT textbooks and identified several dimensions: informational vs. involved production, planned vs. unplanned discourse, narrative vs. non-narrative discourse, non-specific vs. specific reference, linguistic explicitness of rhetoric, and context-dependent vs. context-independent in his corpus.

Van Rooy and Terblanche (2009) tried to distinguish nativelike and non-nativelike uses of English. They used the written production of Tswana learners of English and native speakers of English to develop a new multi-dimensional model that can distinguish style dimensions from grammar and information presentation dimensions that the original model in Biber (1988) did not permit. By creating a new multi-dimensional model, they revealed that grammatical differences have more important implications than writing style differences. They also pointed out three general dimension patterns: a dense informational/nominal structure, a strong oral and informal style, and an intensely persuasive style.

Asención-Delaney and Collentine (2011) used the multi-dimensional approach to show how learners of Spanish use lexical and grammatical features to create different

discourse types when they communicate in writing. They identified four distinguishing discourse types that can be summarized by two stylistic variations, narrative and expository prose. The narrative dimension is characterized by the frequent use of verbal features (e.g., subjunctive, past subjunctive, conditional, progressive aspect) and the expository dimension by nominal features (e.g., nouns, articles, adjectives). Learners used narrative discourse when they described past events as well as personal speculations, feelings, and attitudes toward the events. Learners also used narrative elements even when arguing, a feature of their language use that differs from how native speakers use expository discourse types. The results of the study indicated how stylistic sophistication and linguistic complexity can occur in L2 Spanish writing, but syntactic complexity (e.g., concentrated use of relative clauses, subordinate clauses) and frequent use of nominal features that affect informational density was not found. This was because the data of Asención-Delaney and Collentine (2011) were limited to intermediate high and advanced low learners of Spanish, so Asención-Delaney (2015) used the data of advanced Spanish learners to conduct a multi-dimensional analysis. Asención-Delaney and Collentine (2011) showed how learners depended on limited lexico-grammatical features (i.e., verbal and nominal features) in academic writing (e.g., narration, expository prose, descriptive expository prose, expository prose with stance). However, multiple nominal and verbal features were combined in the expository prose, and a high concentration of nominal features produced a semantically dense and informationally rich text in this study. Some of the same linguistic features which were found in the literate dimension (i.e., nouns, adjectives, definite articles, and type token ratio) were also observed.

Additionally, linguistic features, such as relative clauses were associated with formal discourse in academic prose.

In contrast to the studies, which used Biber's multi-dimensional analysis to gain an overall language description, there is another kind of study, L2 profiling research (Tono, 2013), in the field of learner language research. This type of study is supposed to be connected with language assessment. As pointed out in Chapter 1, Introduction, the Cambridge English Profile Programme has begun to provide criterial features that can describe the proficiency levels of CEFR using data from the Cambridge Learner Corpus (CLC) (Hawkins & Buttery, 2009, 2010; Salamoura & Saville, 2009, 2010). The levels of CEFR were originally described by a statement that was focused on the functional aspect of language, in other words, what learners can do and cannot do by using a target language (e.g., if you can write a short and simple card for a new year's greeting or not). Accordingly, it is useful to describe the proficiency levels in terms of lexical and grammatical aspects. Hawkins and Filipovic (2012) examined how learner language progresses across the levels (from A1 to C2) by tagging and parsing the CLC data. They presented a list of various linguistic features that learners come to use at a certain level and which then persist at all advanced levels (e.g., intransitive and transitive clauses). Even though they found a wide range of linguistic features that can be used as criterial features, it is difficult to know which features have a more important role for distinguishing the levels of Japanese EFL learners.

Consequently, following the trend of using proficiency levels set by CEFR, Negishi (2012) used a checklist of CEFR criterial features to evaluate the written

compositions gathered from 900 Japanese EFL learners, who were assumed to range across 6 levels (A2, A2+, B1, B1+, B2, B2+). The checklist, which includes 53 linguistic features, was drawn from the English Profile Program research with some modifications to suit the Japanese EFL context. He used a program called RASCAL to analyze the data and concluded that 42 criterial features out of the 53 used can effectively discriminate A2 to B2+ level learners. The following features were reported to be especially effective: (a) relative clauses, (b) past/present participial post-nominal clauses, (c) participial constructions, and (d) modal auxiliaries. Additionally, he found that there was no large difference in the number of criterial features that learners used across the A2+, B1, and B1+ levels. However, he also stated that once the learners reach a certain level, in this case B1+, the number of criterial features that learners can use has dramatically increased.

Using a much larger corpus, Alexopoulou, Geertzen, Korhonen, and Meurers (2015) focused on relative pronouns, which are considered to be a difficult linguistic item for Japanese EFL learners. They collected half a million written scripts over a year, in total 33 million words, by using an online language learning platform. The data were collected from 85 thousand learners of 172 nationalities. They belonged to sixteen teaching levels, which can be aligned to CEFR levels. A natural language parser was used for their study, which is trained on native speakers of English, to extract relative pronouns (i.e., *who*, *which*, complementizer *that*), involving 2,369,994 sentences from the five most frequent L1 background learners (i.e., Brazilians, Chinese, Russians, Mexicans, Germans). In Negishi (2012) relative clauses were found to be one of the most effective

features to discriminate the A2 to B2+ levels of Japanese EFL learners, but the data of Japanese learners were not included in this study. They found that learners produced few relative clauses before they reach level 4, and the use increased as the subordinate clauses increased until they reach level 6 (level 4 to level 6 correspond to the CEFR A2 level). Then, after the learners reach level 6, the use stabilizes. In addition to these findings, they showed a strong effect of tasks and national language (in this study they were not able to gain the L1 information) on written production. They claimed that national language had a clear effect on the use of different types of relative clauses.

The research on multi-dimensional analysis and the use of a variety of features across a range of proficiency levels has shown that these can be useful and revealing areas of research. The research has shown the value in looking at a range of language features and the ways in which these features cluster together. The research has also shown how individual language features and clusters of language features can distinguish various L2 proficiency levels that were independently set up on different criteria, typically related to language use and language functions. Moreover, this research has shown that there is likely to be a strong effect of the learners' first language on the use of language features and thus this is an important variable to control or consider when doing such research. If a reasonably standard list of language features is used, such as Biber's (1988) list, then this can allow ready comparison with previous research and allow future research to build on and enrich a growing database of studies.

## **Gaps in the Literature**

The first gap in the literature concerns the lack of studies in which a learner corpus has been used to quantitatively explore variation in interlanguages across a range of oral proficiency levels. Learner language studies involving spoken corpora have been conducted in the past (Abe, 2007a, 2007b, 2008; Crossley, Salsbury, & McNamara, 2010a, 2010b, 2011; Hasselgren, 2002; Housen, 2002; Kobayashi, 2007a, 2007b, 2010). However, the scarcity of spoken learner corpora that contain sufficient proficiency level information has limited the possibility for enriching qualitative learner language studies with computer-based quantitative language analysis. In order to fill this gap, one million tokens of L2 oral performance data coded with English proficiency levels set by the Standard Speaking Test (SST) are used, which range from novice to advanced oral proficiency level. The length of the speaking test, 15 minutes, cannot result in a long individual text size, nonetheless a collection of such texts easily provides enough data for the comparison of oral proficiency. Additionally, the learner corpus used in this study is made up of data gathered from individuals at a single point in time. Thus, it does not employ a longitudinal study, but at least comparing different learners at different oral proficiency levels enables the investigation of variation across oral proficiency levels.

The second gap in the literature is a lack of studies that describe interlanguages using multiple linguistic features. As stated in Biber (1988), “most previous studies of language acquisition/development have focused on a small number of speakers, a single type of language (e.g., narratives), and only a few linguistic features. As a result, the corpus-based approach can make important contributions in this area of study, enabling

comprehensive descriptions of language use at different developmental stages” (p. 180). Some recent studies, such as Van Rooy and Terblanche (2009), Asención-Delaney and Collentine (2011), and Asención-Delaney (2015) have attempted to involve various linguistic features in the study. However, it is still a time-consuming and demanding task for non-technical users to extract frequency information of multiple linguistic items from corpus data. As a result of this gap in the research base, there are few comprehensive descriptions of learner language use. Biber’s (1988) software is not publically available, largely because of the support needed to run it, and researchers have had to develop their own programs or do their research manually.

The third gap in the literature concerns the lack of a research methodology that makes full use of digitized learner performance data to describe the characteristics of learner language. A computer-based approach to learner language can usefully add to SLA data analysis methods and thereby allow researchers to gain a more wide-ranging understanding of spoken learner language. Language processing techniques that can automatically identify targeted linguistic features have not been used to their full potential. The use of a large corpus of learner performance data to look at a wide range of features provides a different and exciting focus for research. Because such research draws on language processing techniques, the frequency information derived from such techniques enables researchers to develop profiles of the variety of language use.

## **The Purposes of this Study and Research Questions**

The first purpose of this study deals the first gap of the study. It aims to identify linguistic features that can be used to describe variation across different oral proficiency groups. As pointed out in the previous section, Abe (2007a) found that particular error types can be markers of a specific English oral proficiency level, but in some cases well-formed linguistic features can more effectively identify English oral proficiency levels (Ellis, 2008). Thus, I proceeded to use the same learner corpus as in Abe (2007a) to focus on well-formed linguistic features. This learner corpus is a collection of oral interview transcripts gathered from test-takers who took part in the Speaking Standard Test (SST). One of major strengths is that each individual script is coded with the learners' oral proficiency level as determined by the SST. Accordingly, this type of learner corpus enables researchers to specify the linguistic features that can describe the characteristics of different oral proficiency groups.

The second purpose of this study is to gain a profile of L2 oral performance from multiple linguistic features. This analysis is unique in calculating the frequency of 58 linguistic features from a spoken learner corpus. In order to deal with such a large number of variables, the NICT JLE corpus was chosen for the analysis. The NICT JLE corpus is particularly suitable because it consists of over one million running words of transcribed spoken performance of more than 1,200 Japanese learners of English, and it is currently considered the largest such corpus (Lüdeling, Kyoto, & McEnery, forthcoming).

The third purpose of this study is to determine whether computer-aided interlanguage analysis of multiple linguistic features can be applied to Japanese EFL

learners' spoken data or not. I examined the effectiveness of using language corpus processing techniques to investigate Japanese EFL learners' oral performance. I used the linguistic features that have been shown to distinguish text and style variation of language in Biber (1988) to examine if they can also be used to distinguish learner language at different oral proficiency levels, and to distinguish learner language from the production of native speakers of English. The learner corpus used in the present study contains more than one million words, but foreign language learners often produce short, simple sentences in their spoken performance that show limited variation in lexical and grammatical use. Thus, in terms of text characteristics it is necessary to verify if the quantitative language analysis method developed for analyzing data produced by native speakers of English can also be used to analyze EFL learners' oral performance.

In order to accomplish these purposes, the following research questions are investigated. The answers to these questions will contribute to understanding the characteristics of learner language from novice to advanced level learners.

1. What linguistic features characterize different English oral proficiency groups of Japanese learners?
2. To what degree do the language features appearing in the spoken production of high proficiency learners match with those of native speakers of English who perform the same task?
3. Is the oral production of Japanese EFL learners rich enough to display the full range of features used by Biber?

I answer each research question by using different statistical analyses, box-and-whisker plots analysis and a multivariate statistical analysis called correspondence analysis.

Box-and-whisker plot analysis examines medians and percentile scores. It shows how the linguistic features are distributed across the oral proficiency groups. Correspondence analysis is useful in dealing with a large number of variables and in summarizing overall tendencies of different oral proficiency groups. It presents linguistic features that can be used to characterize different oral proficiency groups.

## **CHAPTER 3**

### **METHODS**

In this chapter, I provide a description of the quality and quantity of the corpus data and the methodology used in the present study. First, general information about the non-English-speaking learners regarding their (a) gender, age, and social status, (b) overseas experience, and (c) scores on English examinations is provided. Second, the Standard Speaking Test (SST), which the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) corpus is based on, is described in terms of (a) interviewers and raters, (b) English oral proficiency levels, (c) scoring criteria, and (d) assessment tasks. Next, the method of data cleaning used in this study is described and the size of the corpus is shown. Then, the precision rate in frequency profiling, errors and problems in part-of-speech tagging, and error coverage of part-of-speech tagging are discussed. Subsequently, the statistical methods used in this study are explained.

#### **The Learners Who Contributed to the Corpus**

First of all, it is essential provide sufficient information about the Japanese EFL learning context to enable correct interpretation of the results by the readers. I explain the Japanese educational system and learner characteristics in this section. In the Japanese EFL learning context, there is an inadequate amount of exposure to the target language. There is no necessity or opportunity to use English in daily life, and thus English remains

an academic subject and unlikely to become a tool for communication for many learners. Primarily, English is regarded as an instrument for gaining academic success in passing entrance examinations, and the learning of English vocabulary and grammar is typically strongly focused on this limited goal.

Another problem lies in the syllabi and textbooks used in the Japanese ELT context. Different English teachers independently teach each class by using different textbooks. The grammar points are not interconnected with each class, but introduced independently. As a result, learners are required to acquire English grammar not in a collaborative way but in a separate way. What is more, each lesson in ELT textbook is incoherent and it does not include adequate following up tasks to apply newly learned grammar points to real communication.

One more problem worth pointing out is a teaching style based on grammar instruction. A typical English class mainly focuses on the teacher checking the learners' sentence-by-sentence translations of a textbook. The grammar translation method is convenient and comfortable for most teachers, but the classes become heavily teacher centered, and the only task that learners are required to do is to translate the sentences correctly. They do not have enough opportunity to explore or practice using the language during the classes. In short, the application of English knowledge to real communication is not emphasized in the Japanese ELT context, and this results in producing less than competent speakers. Thus, grammar translation-oriented Japanese teaching and the ELT learning context prompted me to examine how the various grammatical features that

should have been learned in English classes are produced across different proficiency levels of oral performance.

The spoken data utilized in this analysis were extracted from the NICT JLE corpus (Izumi, Uchimoto, & Isahara, 2004b), which consists of interview protocol data elicited from a 15-minute oral proficiency test. The data were open to the public in the form of CD-ROM in 2004, and a special website which has open access to the data was accessible in 2012. It is made up of 325 hours of interviews in total conducted with 1,281 Japanese EFL learners. When the corpus was originally created, all the learners who provided data for this corpus were informed that their recorded utterances would be used for research, and only those who agreed to this condition took the test.

The NICT JLE corpus does not provide information on how many years the learners had been studying English. Additionally, information concerning the length of informal English education at institutions such as English conversation schools and cram schools and other foreign languages that they studied were not included in the speaker profiles. However, because the learners were all Japanese born and educated in Japan, it is likely that most of the learners have primarily acquired English in an English as a Foreign Language (EFL) context and their first language is Japanese. As Table 3 shows however, a large proportion of the higher proficiency learners had spent more than a year overseas.

Anyone could take the Standard Speaking Test to find their oral proficiency level so the sample was not restricted by who was allowed to sit the test (now however there is

a 16 year old age restriction but this did not apply to the present study). I investigated the learner data drawing on oral proficiency levels.

### Gender and Age

Table 1 provides general information about the gender and age of learners who belong to SST levels 1-9. Note that the range of gender and the average age of learners who have participated in this speaking test. The total number of learners, the number of males and females, range of age and mean age across SST levels 1-9 are shown. The gender of the learners was almost evenly divided, with 643 males and 638 female learners contributing to the corpus. The number of males was higher than that of females at the novice and intermediate levels (SST levels 1-5), but lower at the higher levels (SST levels 6-9). Of the 1,144 learners who provided their age, the range was from 15 to 70 years old, and the average age was between 25 and 35 across all oral proficiency levels.

*Table 1. Gender and Age Range of the Learners by SST Level*

SST level	1	2	3	4	5	6	7	8	9	Total
Test-takers	3	35	222	482	236	130	77	56	40	1,281
Male	3	23	161	262	91	57	23	16	7	643
Female	0	12	61	220	145	73	54	40	33	638
Age range	21-35	18-65	16-65	16-70	16-65	18-65	16-48	15-69	16-51	-
Mean age	29	26	28	30	31	31	27	29	26	29

*Note.* SST = Standard Speaking Test (SST).

Table 2 provides a more detailed breakdown of ages. All of the test-takers provided their gender, but not all of the test-takers provided their age. Therefore, there are different numbers in some oral proficiency levels in Tables 1 and 2.

Table 2. *The Ages of the Learners by SST Level*

SST Level	1	2	3	4	5	6	7	8	9	Total
11-20	0	14	76	92	27	7	14	15	14	259
21-30	1	11	73	170	91	48	31	14	8	447
31-40	2	5	44	103	48	38	18	16	7	281
41-50	0	3	13	52	24	7	5	5	2	111
51-60	0	0	8	17	8	3	0	0	1	37
61-70	0	1	4	1	1	1	0	1	0	9
Total	3	34	218	435	199	104	68	51	32	1,144

*Note.* SST = Standard Speaking Test (SST).

Overall, there is a good representation of age ranges from teen-agers to those in late middle age. Only 110 learners provided information concerning their social status; 84 were senior high school students, undergraduate, or graduate students.

### **Overseas Experience**

Table 3 shows the overseas experience of the learners who belong to SST levels 1-9. Note the relationship between the amount of time overseas and the learners' English oral proficiency. The learners are divided into four categories of no overseas experience, less than one month, one to twelve months, and more than one year. All of the test-takers provided information on their overseas experience. Five hundred and forty-three learners had not lived overseas, 410 had visited foreign countries for less than one month, 161 had stayed abroad from one month to 12 months, and 167 had lived overseas longer than one year. None of the level 1 and 2 learners had lived overseas more than one month, whereas over almost half of the learners in levels 7 to 9 had lived overseas for more than one year. While we would expect that as the amount of time overseas increased, the learners'

English oral proficiency would also increase, it is interesting to note that there were very high proficiency level learners (Level 9) with no overseas experience, and intermediate level learners with substantial overseas experience. Not all benefit in the same way from overseas experience.

Table 3. *Overseas Experience of the Learners by SST Level*

SST level	1	2	3	4	5	6	7	8	9	Total
No experience	3	25	112	203	97	49	22	16	16	543
Less than 1 month	0	10	96	203	71	21	5	4	0	410
1 to 12 months	0	0	9	53	45	37	13	3	1	161
More than 1 year	0	0	5	23	23	23	37	33	23	167

*Note.* SST = Standard Speaking Test (SST)

### **Scores on English Examinations**

Table 4 shows the number of learners who have scores from other English examinations, TOEIC and TOEFL. It also shows the means and standard deviations of learners who belong to SST levels 1-9. Information on scores on other English examinations is valuable in understanding the general English proficiency level of the non-English-speaking learners and for checking the effectiveness of the SST. TOEIC scores were reported by 565 learners and TOEFL scores were reported by 97 learners. The TOEIC scores ranged from 356 to 921 with a mean of 690. The TOEFL scores ranged from 442 to 608 and the mean score was 545. To allow a comparison between learners providing different types of TOEFL scores, all computer-based test scores were converted into paper-based test scores using the TOEFL Internet-based test score comparison tables (ETS, 2005). The relatively wide range in both the TOEIC and TOEFL

scores confirms that the learner data contain a wide range of proficiency levels, and there is a consistent increase in TOEIC and TOEFL mean scores as the SST level increases.

This supports the idea that the SST levels are measuring language proficiency.

## **Instrumentation**

### **Interviewers and Raters**

One of the advantages of using the NICT JLE corpus is that all the non-English-speaking learners are assigned to one of nine English oral proficiency levels based on their performance on the SST. The SST is an interactive one-on-one speaking test, which is adapted to the perceived oral proficiency level of the test-takers.

Interviewers and raters are required to participate in a series of certification workshops organized by ALC Press, the publisher that developed this oral English proficiency test with ACTFL, and both interviewers and raters are required to pass the examination based on the workshops. Interviewers are trained to adapt to the perceived oral proficiency level and personal and professional interests of test-takers to elicit performances that ensure the most accurate rating. They are also required to practice changing topics and asking different types of questions to determine which oral proficiency level test-takers can maintain and which they fail to sustain. The interviewers, however, do not formally evaluate the oral proficiency of the test-takers. Two raters evaluate the recorded oral performance drawing on the SST scoring criteria that are not publically available but are summarized in Appendix A using the information retrieved from the ALC Press Web site

Table 4. Scores on the TOEIC and TOEFL Examinations by the Test-Takers

SST level	1	2	3	4	5	6	7	8	9	All levels
Number of TOEIC takers	0	9	71	224	116	66	36	28	15	565
TOEIC mean score	—	355.56	494.23	641.71	730.69	805.65	877.08	868.04	921.00	689.66
TOEIC S.D.	—	47.79	132.95	101.77	85.98	88.50	69.51	77.89	53.49	—
Number of TOEFL takers	0	0	2	19	25	19	14	9	9	97
TOEFL mean score	—	—	441.50	510.05	528.80	538.61	570.96	600.28	607.89	545.30
TOEFL S.D.	—	—	143.54	42.78	46.68	34.00	39.76	24.47	52.20	—

Notes. SST = Standard Speaking Test (SST). To allow a comparison between learners providing different types of TOEFL scores, all computer-based test scores were converted into paper-based test scores using the TOEFL Internet-based test score comparison tables (ETS, 2005).

(<http://www.alc.co.jp/edusys/sst/e/index.html>). According to this Web site, both the first and the second rater independently rate the recorded oral performance. In cases in which the two raters award different ratings, a third rater, the master rater, finalizes the rating.

The grammatical criteria in the rating scale are phrased in very general terms (e.g., *mostly correct simple sentences; compound sentences; good command in using tenses*) and typically in terms of error (e.g., *omission of conjunctions; elementary errors sometimes occur in complex sentences but not habitual; can produce grammatically correct speech unconsciously*). In addition, some of the grammatical features mentioned under the text type heading in the criteria occur at more than one of the proficiency levels (e.g., *simple sentences, compound sentences*). None of the grammar features in the criteria, except for present tense and past tense, directly overlap with the features investigated in this study. Thus there is no risk that the present study is circular in that the criteria used to decide oral proficiency levels include the same features that are investigated in this study. It is also worth noting that although there are criteria to guide raters' judgments, in essence the rating is done holistically because many of the criteria are rather vague (e.g., *lack of basic knowledge; frequent minor errors*). I took the four-day interviewer and rater training session and used to be an SST interviewer. Both native and non-native speakers of English can be interviewers, but it is necessary to have an ACTFUL OPI advanced level certification and to pass a test to become an interviewer.

### **English Oral Proficiency Levels**

There are nine levels in the SST, level 1 (novice low) to level 9 (advanced) proficiency levels. The test was designed to be used with Japanese EFL learners, who

generally have low oral proficiency in English. However, the evaluation criteria conform to those used with the American Council on the Teaching of Foreign Language Oral Proficiency Interview (ACTFL OPI). The novice level on the SST is subdivided into three ranks (low, mid, and high) as in the ACTFL OPI, but the intermediate level is divided into five ranks (low, low-plus, mid, mid-plus, and high) to discriminate Japanese learners of English, who mostly belong to this level. Furthermore, unlike the ACTFL OPI, which has three ranks for the advanced level (advanced low, advanced mid, and advanced high), the SST has only one rank to categorize advanced proficiency learners; thus, the SST is not designed to subdivide test-takers who are at an advanced proficiency level. Table 5 compares how oral proficiency levels are differently categorized in the SST and ACTFL OPI.

Table 5. *A Comparison of the Levels of the SST and ACTFL OPI (Retrieved July 11, 2011 from ALC Press Website: <http://www.alc.co.jp/edusys/sst/e/index.html>)*

ACTFL OPI levels	SST levels
Superior	Level 9 (Advanced)
Advanced High	
Advanced Mid	
Advanced Low	
Intermediate High	Level 8 (Intermediate High)
Intermediate Mid	Level 7 (Intermediate Mid-plus)
	Level 6 (Intermediate Mid)
Intermediate Low	Level 5 (Intermediate Low-plus)
	Level 4 (Intermediate Low)
Novice High	Level 3 (Novice High)
Novice Mid	Level 2 (Novice Mid)
Novice Low	Level 1 (Novice Low)

*Note.* SST = Standard Speaking Test (SST).

Figure 1 shows the number of learners at each oral proficiency level (Level 1: 3, Level 2: 35, Level 3: 222, Level 4: 482, Level 5: 236, Level 6: 130, Level 7: 77, Level 8: 56, Level 9: 40). The number of learners in level 4 (intermediate low) is

considerably higher than the number at the other levels, while the numbers of learners in levels 1, 2, and 9 are low.

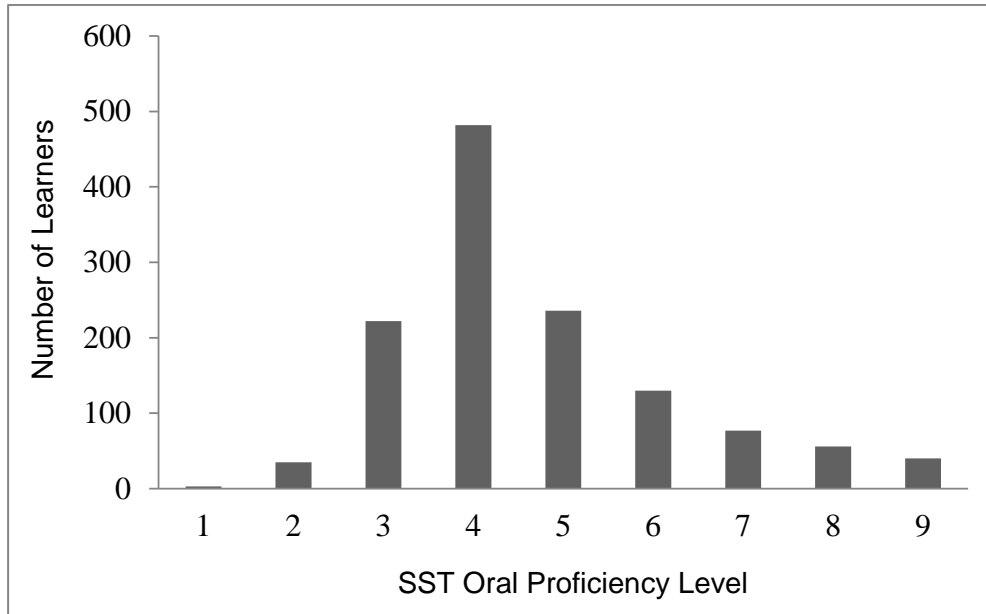


Figure 1. Number of learners at each SST oral proficiency level ( $N = 1,281$ ).

### Scoring Criteria

Raters score the test-takers using four assessment categories: (a) global tasks and functions, (b) social context and content area, (c) accuracy of grammar, vocabulary, pronunciation and fluency, and (d) text types and quantity of speech. The descriptions summarize how English proficiency levels are discriminated under these criteria (Appendix A).

### Assessment Tasks

The SST takes approximately 15 minutes to administer, test-takers are not provided with any planning time, and the use of reference material is not allowed. The

test is divided into the following five stages

(<http://www.alc.co.jp/edusys/sst/interview.html>):

Stage 1: Warm-up questions for initial assessment (3-4 minutes)

Stage 2: A single picture description task and questions (2-3 minutes)

Stage 3: A role-play task and questions (1-4 minutes)

Stage 4: A narrative task with a 4 or 6 picture sequence (2-3 minutes)

Stage 5: Wind-down questions (1-2 minutes)

At Stage 1, the interviewer asks simple questions (e.g., How are you doing today?) to break the ice and to probe the proficiency level of the test-taker. Then, at Stage 2 the test-taker is shown a picture and asked to describe it and answer questions related to the picture or the content of the utterance the test-taker has made. This single picture description task is not designed to elicit specific grammatical points or functional expressions but to extract natural and spontaneous language use. The interviewer evaluates the initial Stage 2 assessment based on how the test-taker uses the present tense to describe the picture. At Stage 3, the test-taker plays a role in a specific context such as at the train station or store. This role-play task is a communicative language activity replicating a real life situation to determine whether the test-taker can use particular functional expressions. The test-takers' ability to ask questions, make requests, or give reasons appropriately, is checked. At Stage 4, the test-takers are shown a sequence of four or six pictures and asked to tell the story and then answer questions related to the picture. This narrative picture sequence task is used to determine whether the examinees can properly use the past and present tenses to narrate a story. Finally, at Stage 5, easy questions are asked to reduce any tension associated with the speaking test. All the stages of examination were used for the analysis in the present study.

The data were gathered during an English speaking test in which the test-takers did not have preparation time. Accordingly, the test-takers might have paid much more attention to grammatical accuracy than grammatical complexity of utterances to gain a better test result. They might have used simple sentences and phrases to avoid making errors that can cause misunderstandings. If this is the case, the test-takers' strategies might have reduced the complexity of the resulting oral performance. However, the results of this study indicated that the complexity of sentence structures increased as the oral proficiency level rose. Therefore, there is some justification for discussing the oral production of the learners gathered from the oral proficiency test as a product that is similar to a natural performance.

## **Corpus Data**

### **Corpus Construction**

Before explaining the data cleaning method, the corpus construction methodology of Izumi, Uchimoto, and Isahara (2004b) is summarized briefly. Under the NICT's leadership, after the recorded spoken data were transcribed manually, basic discourse tags were inserted and the transcriptions were proofread and corrected. If transcribers were able to understand words mispronounced by the learners (e.g., *right* and *light*) from the context, they were corrected. However, when transcribers encountered a problem identifying a word, they indicated this with single-question-mark tags (<?>...</?>); when utterances were not clear enough to understand, double-question-mark tags (<??></??>) were inserted. Repetition tags (<R>...</R>) were used when learners repeated the same word or expression, and these tags were added into the first utterance rather than following utterances. Also, question marks were added to tags such as (<R?>...</R?>) whenever utterances were

not clear enough to understand. Additionally, utterances rephrased or corrected by speakers were indicated by self-correction tags (<SC>...</SC>, <SC?>...</SC?>). These tags were added until test-takers stopped making corrections. Table 6 summarizes the discourse tags used in this learner corpus. The information is extracted from the basic discourse tags guideline version 2.1.3 (Izumi, Uchimoto, & Isahara, 2004b).

Table 6. *Basic Discourse Tags Guideline ver. 2.1.3 (extracted from Izumi, Uchimoto, & Isahara, 2004b)*

Discourse tag	Function
<F></F>	<b>F</b> iller / <b>F</b> illed pause
<R></R>	<b>R</b> epetition
<R?></R?>	<b>R</b> epetition? (lack of confidence in understanding utterance)
<SC></SC>	<b>S</b> elf- <b>C</b> orrection
<SC?></SC?>	<b>S</b> elf- <b>C</b> orrection? (lack of confidence in understanding utterance)
<CO></CO>	<b>C</b> ut <b>O</b> ff
<?></?>	lack confidence to transcribe
<??></??>	impossible to transcribe
<H pn="X"></H>	<b>H</b> idden (e.g., proper noun, discriminatory words)
<JP></JP>	<b>J</b> a <b>P</b> anese
<.></.>	short pause (2 to 3 seconds)
<..></..>	long pause (longer than 3 seconds)
<OL></OL>	<b>O</b> ver <b>L</b> apping
<nvs></nvs>	<b>N</b> on- <b>V</b> erbal <b>S</b> ound
<laughter></laughter>	<b>L</b> aughter
<ctxt></ctxt>	<b>C</b> on <b>T</b> e <b>X</b> T

In addition to the information on discourse tags in Table 6, Table 7 show how these tags were inserted into the transcribed spoken data. The examples of basic discourse tags are from Izumi, Uchimoto and Isahara (2004b).

Table 7. *Examples of Basic Discourse Tags* (Izumi, Uchimoto, & Isahara, 2004b)

Tag	Example
<F></F>	<F>ah</F> <F>mm</F> <F>mhm</F> <JP><F>etto</F></JP>
<R></R>	When <R>he</R> <R>he</R> he was a child...
<SC></SC>	He <SC>don't</SC> doesn't know anything about this.
<CO></CO>	<A><F>Oh</F> O K. So it's getting dark but is it O K for you to come out? <CO>Is that</CO>.</A>
<?></?>	<?>They</?> should be very beautiful.
<??></??>	<??></??> should be very beautiful.
<H pn="X"></H>	<A><H pn="A's name">Hanako Yamada</H>. May I have your name?</A>
<JP></JP>	<F>Mm</F> <R>I</R> I don't like <JP>osechi</JP>.
<OL></OL>	<A>So what are you going to do <OL>this weekend</OL>?</A> <B><OL><F>Oh</F> yes</OL> That's what I'm going to tell you about.</B>
<nvs></nvs>	<nvs>laughter</nvs>, <nvs>sigh</nvs>, <nvs>cough</nvs>, <nvs>yawn</nvs>
<laughter></laughter>	It's a kind of <laughter><JP>mama-chari</JP></laughter>.
<ctxt></ctxt>	<ctxt>The interviewer is choosing a card.</ctxt>

## Reference Corpus

In addition to the NICT JLE corpus, a reference corpus was constructed (Izumi, Uchimoto, & Isahara, 2004b). The normative data consist of 20 native speakers of English who performed the same speaking tasks as the learners, which consist of 84, 774 words. The native speakers of English were interviewed by official interviewers using official materials, so that the conditions for their test were similar to the conditions learners were tested under. However, unfortunately, information about age, nationality, gender and educational level of these native speakers of English is not available. It would be ideal to have more data about the native speakers of English and it is necessary to be cautious when using such a small amount of data when making comparisons with foreign language learner use. However, these data can be used to investigate which vocabulary or grammatical structures the English

speakers used frequently or infrequently in the interviews. The strength of the data is that the native speakers of English performed the same tasks as the non-native speakers of English.

### **Data Cleaning**

The data cleaning method involved the following steps. The present study is focused on how learners use the target language in oral performance; therefore, marked up information regarding overlaps (<OL>...</OL>), cut offs (<CO>...</CO>), and Japanese (<JP>...</JP>) was not deleted as it is related to learners' language use. Some utterances and annotations that were not targeted in this analysis were deleted from the database and not included in the total number of tokens. Leaving them in could have upset frequency figures and the recognition of the 58 targeted features. The following utterances and annotations: repeated speech, self-correction procedures, marginal words (short pauses and longer pauses), fillers, non-verbal sounds and laughter were removed from the data.

- repeated speech (<R>...</R>, <R?>...</R?>)
- utterances prior to self-correction (<SC>...</SC>, <SC?>...</SC?>)
- short pauses that last for two to three seconds (<.></.>)
- longer pauses that last more than three seconds (<..></..>)
- fillers (<F>...</F>)
- non-verbal sounds (<nvs>...</nvs>)
- laughter (<laughter></laughter>)
- context (<ctxt>...</ctxt>)

Quotation marks were removed from the data, because they can affect the accuracy rate of the part-of-speech (POS) tagging. Unnecessary periods and spaces between

upper-case letters were also removed from the data by manually checking the corpus wordlist. Table 8 shows examples of unnecessary periods and spaces between letters, which were removed from the texts. These examples of before displacement and after displacement can show how unnecessary periods and spaces were taken out from the texts. This enabled me to obtain accurate total word counts and to increase the precision of the automatic POS tagging and counting of linguistic features. In some corpus-based studies, contracted forms are modified to achieve the same purpose, but as the computer program used in this analysis can count contracted forms (e.g., *I'm*, *it's*, and *can't*) as one word and divide them into two words for the POS tagging, no modification was made.

**Table 8. *Unnecessary Periods and Spaces Removed from the Texts***

---

Examples of unnecessary periods between letters:
Before displacement: O.K / O.K, / O.K. / O.K?
After displacement: OK / OK, / OK. / OK?
Examples of unnecessary spaces between letters:
Before displacement: B B C / O K / C D / U C L A / T V
After displacement: BBC / OK / CD / UCLA / TV

---

The original data shown below include the utterances produced by an interviewer (<A>...</A>) and interviewee (<B>...</B>). They include unnecessary information for the analysis.

<A>Hello. My name is <H pn="A's name">XXX01</H>. What is your name?</A>

<B>My name is <H pn="B's name">XXX02</H>.</B>

<A>Hello, <H pn="B's name">XXX02</H>.</A>

<B><OL>Hello</OL>.</B>

<A><OL>How are you</OL>?</A>

<B><F>Um</F> I'm fine, but I'm a little <laughter>nervous</laughter>.</B>

<A><F>Ah</F> it's O K.</A>

<B><nvs>laughter</nvs></B>

<A><F>Mhm</F> so, have you been busy these days?</A>

<B>Yes. <F>Mm</F> because, <F>mmm</F> <R>I</R> I had to hand in my graduation thesis in December,</B>

<A><F>Uhm</F>.</A>

<B>so <F>mmm</F> <R>I</R> <F>mm</F> I was busy, but <F>mm</F> now I finished it. <CO>So</CO>. <OL><F>Mmm</F></OL>.</B>

<A><OL><F>Oh</F> good</OL>.</A>

<B>Yes. <F>Um</F>.</B>

<A><F>Uhm</F> I see. So now, <F>err</F> winter vacation is soon coming, right?</A>

<B><F>Mm</F> yes, but now, not yet <F>mm</F>. <F>Mmmm</F> maybe next week, <F>mm</F> winter vacation <F>mm</F> will start.</B>

<A>I see. What's you plan for the holiday?</A>

<B><nvs>laughter</nvs> I don't have any plan. <nvs>laughter</nvs>

<F>Uhm</F> maybe, <F>mmm</F> <F>mmm</F> <SC>I work</SC>

<F>mm</F> <R>I</R> <F>um</F> I work for five or six days.

The data shown below is the same sample as above but has been cleaned-up for the present study.

My name is XXX02.

Hello.

I'm fine, but I'm a little nervous.

Yes.

because, I had to hand in my graduation thesis in December, so I was busy, but now I finished it.

So.

Yes.

yes, but now, not yet.

maybe next week, winter vacation will start.

I don't have any plan.

maybe, I work for five or six days.

### **Counting Overall Corpus Size**

The total number of non-English-speaking learners initially was 1,281 including three learners in level 1 and 35 learners in level 2. The 38 level 1 and level 2 learners were deleted from the data because an interviewer does most of the speaking during the interview and the test-takers quite possibly repeated the words or phrases used by the interviewer. These two SST proficiency levels are described on the website (<http://www.alc.co.jp/edusys/sst/e/index.html>) as follows: test-takers at these oral proficiency levels have difficulty producing sentences, so they use Japanized English words or short memorized phrases repeatedly between long pauses, and many of their sentences are grammatically incorrect and incomplete.

After deleting level 1 and level 2 files, the total data size of corpus was calculated. As different concordancing programs do not define words and letters identically, the total numbers of words and mean word length differ. Word forms are usually recognized as a sequence of alphabetic or alphanumeric characters that are not interrupted by whitespace, such as spaces, tabs, and newlines (Gries, 2008). However, the differences in numbers are caused by how programs recognize the following: (a)

punctuation (e.g., I'll / e.g. / etc.), (b) hyphenation (e.g., twenty-five / e-mail), (c) numbers (e.g., 2009, XXX02, 25th), and (d) tags (e.g., <head></head>).

Consequently, it is crucial to examine how the program counts words. In this study, a script written in the computer programming language Perl counted the following items as one word: (a) contractions, (b) ordinal and cardinal numbers, and (c) words connected by periods. For example, "2009" and "2010" were treated as two separate words, and hyphenated words were not counted as one word. Therefore, the total token count in the following example is 12 (I'd like 25th December 2009 two thousand nine, and e.g. twenty-five.).

Table 9 shows the descriptive statistics for each oral proficiency level. It shows the number of test-takers, tokens, word types, and Guiraud index. As the table shows, the most striking distinguishing feature of the proficiency levels is the average length of the spoken text, that is how much learners speak and probably how fluently they speak, with native speakers of English saying ten times more than L3 learners in the same time, and three times as much as L9 learners. The average text length can be a measure of oral proficiency, and the increase of text length shows that the higher proficiency level is characterized by greater fluency.

In addition to these descriptive statistics, Figure 2 shows the differences of mean tokens across the oral proficiency levels. In total 1,243 non-English-speaking learners participated in the present study along with 20 native speakers of English making a total of 1,263 test-takers, but as the number of learners differed considerably from one level to another, the file sizes of the oral proficiency levels differed; the smallest had 54,394 words (level 9) and the largest 308,544 words (level 4).

Table 9. *Descriptive Statistics for the SST Oral Proficiency Levels*

	3	4	5	6	7	8	9	NS	Total
Test-takers	222	482	236	130	77	56	40	20	1,263
Total tokens	95,352	308,544	204,048	130,678	85,395	68,539	54,394	84,774	1,031,724
Minima	181	360	498	580	745	865	825	2983	181
Maxima	884	1,151	1,479	1,587	1,701	2,111	2,019	5,840	5,840
<i>M</i>	429.51	640.13	864.61	1,005.22	1,109.03	1,223.91	1,359.85	4,238.70	816.88
<i>SD</i>	144.25	28.28	469.52	103.24	141.42	771.45	13.44	469.52	—
Total word types	47,843	136,665	83,026	51,035	32,517	25,361	19,529	22,810	418,786
Minima	109	186	229	287	311	323	352	915	109
Maxima	337	457	496	527	653	654	683	1,380	1,380
<i>M</i>	215.51	283.54	351.81	392.58	422.30	452.88	488.23	1140.50	331.58
<i>SD</i>	36.06	1.41	164.05	25.46	39.60	185.26	5.66	136.47	—
Guiraud index	25.87	24.33	23.8	24.44	23.14	23.28	22.33	26.42	—

*Notes.* Level 3 is the lowest proficiency level used in this study. As the text length in each file varies, an adapted measure, the Guiraud index, was used to estimate lexical richness. It is calculated by dividing the number of tokens by the square root out of the total number of different word tokens, types/ $\sqrt{\text{tokens}}$ . This formula is designed to measure lexical diversity of texts with different lengths (Jarvis, 2002) and this measurement is also suitable for measuring learner data (Daller, van Hout, & Treffers-Daller, 2003). NS = Native speaker of English.

However, as shown in Figure 2, the mean word count across oral proficiency levels increased as learners' proficiency level increased and a huge gap only existed between that of native and non-native speakers of English.

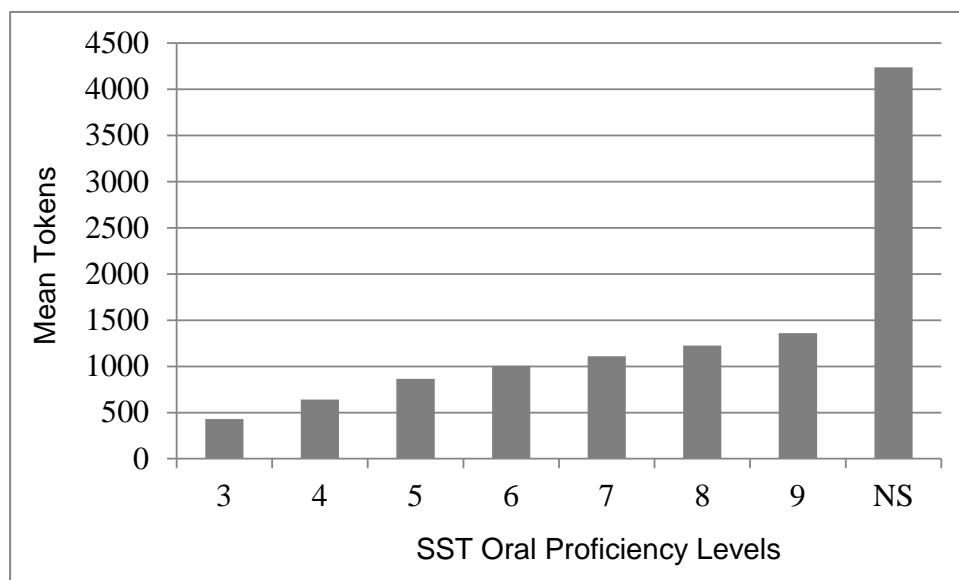


Figure 2. Mean tokens for the SST oral proficiency levels ( $N = 1,263$ ). NS = Native speaker of English.

The speaking test length, 15 minutes, used in this study does not result in a long individual text size, but there is a clear difference in length (number of tokens) from one proficiency level to another. The higher the level, the greater the length. The learner corpus used in the present study is not made up of data gathered from the same individuals over an extended period of time. Thus, this study cannot be viewed as a longitudinal study. Many learner language studies are cross-sectional, and they provide a limited view of how language changes and develops from one proficiency level to another during the process of language learning. However, at least, the NICT JLE corpus is considered the largest among the spoken learner corpora up to the present time (Lüdeling, Kyto & McEnery, forthcoming).

## Linguistic Features

In the present study, 58 linguistic features were used from the original list of 67 linguistic features in Biber (1988). The features are classified into following major categories: tense and aspect markers, place and time adverbials, pronouns and pro-verbs, questions, nominal forms, passives, stative forms, subordination, prepositional phrases, adjectives, and adverbs, lexical classes, modals, specialized verb classes, reduced forms and dispreferred structures, coordination, and negation. The following seven items: (a) demonstratives, (b) gerunds, (c) present participial clauses, (d) past participial clauses, (e) present participial WHIZ deletion relatives, (f) sentence relatives, and (g) subordinator-*that* deletion could not be included in the analysis. This is mainly because of a difference in the software used to annotate part-of-speech tags compared with Biber, and programming problems in automatically extracting these linguistic features. In addition, type-token ratio and word length were excluded because they involve lexical rather than grammatical information. Table 10 shows the 58 linguistic items that are analyzed in this study. I use the same names of linguistic categories that were used in Biber (1988).

As Biber (1988) aimed to analyze linguistic variation between the spoken and written production of native speakers of English, he did not deal with linguistic features that might characterize foreign language learners. For example, morphological features, such as derived words functioning as verbs and articles were not included in the list. Thus, there is a problem in Biber's (1988) list of grammatical features in terms of not dealing with linguistic features that might characterize foreign language learners. In addition, vocabulary features, such as phrasal verbs, word n-grams, or part-of-speech n-grams were not included in this list. An n-gram is a

Table 10. *Fifty-Eight Linguistic Features Analyzed in the Present Study*

---

- A. Tense and aspect markers
  - 1. past tense
  - 2. perfect aspect
  - 3. present tense
- B. Place and time adverbials
  - 4. place adverbials (e.g., *across, behind, inside*)
  - 5. time adverbials (e.g., *early, recently, soon*)
- C. Pronouns and pro-verbs
  - C1. Personal pronouns
    - 6. first person pronouns
    - 7. second person pronouns
    - 8. third person pronouns (excluding *it*)
  - C2. Impersonal pronouns
    - 9. pronoun *it*
    - 10. demonstrative pronouns (*that, this, these, those*)
    - 11. indefinite pronouns (e.g., *anybody, nothing, someone*)
  - C3. Pro-verbs
    - 12. pro-verb *do* (e.g., *the cat did it*)
- D. Questions
  - 13. direct WH-questions
- E. Nominal forms
  - 14. nominalizations (ending in *-tion, -ment, -ness, -ity*)
  - 15. total other nouns (except for nominalizations)
- F. Passives
  - 16. agentless passives
  - 17. *by*-passives
- G. Stative forms
  - 18. *be* as main verb
  - 19. existential *there* (e.g., *there are several explanations . . .*)
- H. Subordination
  - H1. Complementation
    - 20. *that* verb complements (e.g., *I said that he went*)
    - 21. *that* adjective complements (e.g., *I'm glad that you like it*)
    - 22. WH-clauses (e.g., *I believed what he told me*)
    - 23. infinitives (*to*-clause)
  - H2. Participial forms
    - 24. past participial postnominal (reduced relative) clauses (e.g., *the solution produced by this process*)
  - H3. Relatives
    - 25. *that* relatives in subject position (e.g., *the dog that bit me*)
    - 26. *that* relatives in object position (e.g., *the dog that I saw*)
    - 27. WH relatives in subject position (e.g., *the man who likes popcorn*)
    - 28. WH relatives in object position (e.g., *the man who Sally likes*)
    - 29. WH relatives with fronted preposition (e.g., *the manner in which he was told*)
  - H4. Adverbial clauses
    - 30. causative adverbial subordinators: *because*
    - 31. concessive adverbial subordinators: *although, though*
    - 32. conditional adverbial subordinators: *if, unless*
    - 33. other adverbial subordinators: (having multiple functions) (e.g., *since, while, whereas*)
- I. Prepositional phrases, adjectives, and adverbs
  - 11. Prepositional phrases
    - 34. total prepositional phrases

---

(Table 10 continues).

(Table 10 continued).

- 
- I2. Adjectives and adverbs
    - 35. attributive adjectives (e.g., *the big horse*)
    - 36. predicative adjectives (e.g., *the horse is big*)
    - 37. total adverbs (except conjuncts, hedges, emphatics, discourse particles, downtoners, amplifiers)
  - J. Lexical classes
    - 38. conjuncts (e.g., *consequently, furthermore, however*)
    - 39. downtoners (e.g., *barely, nearly, slightly*)
    - 40. hedges (e.g., *at about, something like, almost*)
    - 41. amplifiers (e.g., *absolutely, extremely, perfectly*)
    - 42. emphatics (e.g., *a lot, for sure, really*)
    - 43. discourse particles (e.g., sentence initial *well, now, anyway*)
  - K. Modals
    - 44. possibility modals (*can, may, might, could*)
    - 45. necessity modals (*ought, should, must*)
    - 46. predictive modals (*will, would, shall*)
  - L. Specialized verb classes
    - 47. public verbs (e.g., *acknowledge, admit, agree*)
    - 48. private verbs (e.g., *anticipate, assume, believe*)
    - 49. suasive verbs (e.g., *agree, arrange, ask*)
    - 50. *seem* and *appear*
  - M. Reduced forms and dispreferred structures
    - 51. contractions
    - 52. stranded prepositions (e.g., *the candidate that I was thinking of*)
    - 53. split infinitives (e.g., *he wants to convincingly prove that ...*)
    - 54. split auxiliaries (e.g., *they are objectively shown to ...*)
  - N. Coordination
    - 55. phrasal coordination (e.g., NOUN *and* NOUN, ADJ *and* ADJ)
    - 56. independent clause coordination (clause initial *and*) (e.g., *It was my birthday and I was excited.*)
  - O. Negation
    - 57. synthetic negation (e.g., *no answer is good enough for Jones*)
    - 58. analytic negation: *not* (e.g., *that's not likely*)
- 

Note. Some of the examples were taken from Biber and Conrad (2009).

sequence of words of a given length (n) where the words occur immediately following each other, so, *as a matter of* is a four item n-gram. Note that n-grams need not be grammatically complete units but are simply defined by their length in number of words. As introduced in Chapter 2, Review of the Literature, Aarts and Granger (1998) investigated the sequential use of parts-of-speech by language learners, and revealed similar overuse and underuse tendencies among Dutch, Finnish, and French learners. Tono (2000) also focused on the written production of Japanese learners of English by analyzing part-of-speech tag sequences, and found that these develop in

the order of nouns, verbs, and prepositional phrases. However, there were no clear developmental indices of n-grams and part-of-speech n-grams that can clearly discriminate oral proficiency levels. To make such a list, it might be necessary to investigate specific grammatically-well-structured multiword units such as *as a result*, *at the end of the day*, *on the other hand* in a future learner corpus study.

The idea of using the same linguistic feature set as Biber did for analysis in this thesis is important, because the results can then be compared with other similar studies, for example, Biber's (2006) study of language use in universities (e.g., textbooks, academic speech), and Murakami's (2009) study of Asian ELT textbooks. The present study usefully adds to these previous studies, particularly as Asención-Delaney (2015) pointed out that "the MD analysis would have benefited [from] the inclusion of learners' oral samples (e.g., interviews, presentation) to understand academic L2 discourse where there are high real-time (i.e. cognitive) demands" (p. 264). Consequently, I used the advantage of applying the same list to examine the linguistic variation of Japanese EFL learners' oral performance. In addition, it is worth mentioning that Biber took great care in developing his list, drawing extensively on previous research on genre and text differences, and it has not yet been replaced by a superior list. It is an inclusive and well researched list, and is detailed enough to provide useful insights into the oral proficiency levels of foreign language learners.

Table 11 shows how the various grammatical items from the research I discussed in literature review that dealt with syntactic items overlapped with those that were included by Biber and consequently in this study. This is an additional confirmation of the value of Biber's list.

Table 11. *Grammatical Features from the Previous Studies that Overlapped with Those Analyzed in the Present Study*

Study	Features	Mode	Learners	Findings
Tono (2000)	POS sequence	written	Japanese	<ul style="list-style-type: none"> <li>Learners at the beginning level use verb-related trigrams in high frequency.</li> <li>Learners at lower-intermediate and advanced level use noun-related trigrams in high frequency.</li> <li>Learners at the highest level use preposition-related trigrams in a high frequency.</li> <li>Article- and modal auxiliary-related trigrams were fairly less frequently used, and they can be used as discriminatory features to distinguish lower and higher school-year groups.</li> <li>The writing of Japanese EFL learners develops in the order of noun-, verb-, and preposition-related POS tag sequences.</li> </ul>
Granger & Rayson (1998)	word category	written	Advanced French	<ul style="list-style-type: none"> <li>Overused features by learners are: (a) indefinite articles, (b) most indefinite determiners, (c) most indefinite pronouns, (d) coordinating conjunctions but and or, (e) some complex subordinators, (f) short adverbs of native origin used for place and time, (g) auxiliaries, and (h) infinitives.</li> <li>Underused features by learners are: (a) definite articles, (b) the coordinating conjunction "and," (c) most subordinators, (d) most prepositions, (e) most adverbial particles, (f) -ly adverbs, (g) nouns, and (h) -ing and -ed participles.</li> </ul>
Kobayashi (2007b); Housen (2002)	word category verb tense	spoken	Japanese Dutch French	<ul style="list-style-type: none"> <li>The spoken performance of Japanese EFL learners develops in the order of nouns to verbs.</li> <li>There was a significant variation among individual learners in using verbal forms.</li> </ul>
Lorenz (1998)	adjective intensifiers	written	German	<ul style="list-style-type: none"> <li>Learners overused intensified adjectives (e.g., important, good, successful).</li> <li>The frequent use of intensified adjectives indicates the linguistic immaturity and non-native-likeness of the learners.</li> <li>The number of attributive adjectives correlates with linguistic maturity in the writing of native speakers of English, but not for learners.</li> <li>Both native and non-native speakers of English use more predicative adjectives than attributive intensification.</li> <li>Learners attempted to pack too much information into their sentences by using adjective intensification in thematic places.</li> </ul>
Aijmer (2002)	modals	written	French German Swedish	<ul style="list-style-type: none"> <li>The overuse of modal auxiliaries, modal adverbials, and lexical verbs with modal meaning was a shared characteristic of L2 writers, and partly reflected developmental and interlingual characteristics of learner language use.</li> </ul>

(Table 11 continues).

(Table 11 continued).

Study	Features	Mode	Learners	Findings
Vyatkina (2013)	syntactic complexity	written	German	<ul style="list-style-type: none"><li>• There is a general developmental trend in the increase of the frequency and range of syntactic complexity features (e.g., coordinated, nominal, and nonfinite verb structures).</li></ul>
Haiyang & Lu (2015)	syntactic complexity	written	Chinese	<ul style="list-style-type: none"><li>• Most of ten syntactic complexity measures (e.g., unit length, amount of subordination and coordination, degree of phrasal sophistication) regarding subordination showed significant differences between different learners' proficiency levels.</li><li>• The degree of coordination did not show differences.</li></ul>
Negishi (2012)	CEFR criteria features	written	Japanese	<ul style="list-style-type: none"><li>• 42 criterial features can effectively discriminate A2 to B2+ level learners.</li><li>• The following features are especially effective: (a) relative clauses, (b) past/present participial post-nominal clauses, (c) participial constructions, and (d) modal auxiliaries.</li><li>• There was no large difference in the number of criterial features that learners can use across the A2+, B1, and B1+ level.</li><li>• Once the learners reach a certain level, in this case B1+, the number of criterial features that they can use has drastically increased.</li></ul>
Alexopoulou, Geertzen, Korhonen, & Meurers (2015)	relative clauses	written	Brazilians Chinese Russians Mexicans Germans	<ul style="list-style-type: none"><li>• Learners produce few relative clauses before they reach level 4, and the use increase as the subordinate clauses increase until they reach level 6 (level 4 to level 6 correspond to the CEFR A2 level).</li><li>• Then, after the learners reach level 6, the use stabilizes.</li><li>• National language has a clear effect on different types of relative clauses.</li></ul>
Abe (2007a)	multiple errors	spoken	Japanese	<ul style="list-style-type: none"><li>• There are some categories (e.g., voice, modal verb, finite and nonfinite verb, verbal form, verb complement) that cannot be simply explained by the patterns of what learners do wrongly but can be possibly more clearly explained by what they do correctly.</li></ul>

## Part-of-Speech Tagging

Both untagged and part-of-speech tagged texts were used to count the frequencies of the targeted linguistic features. The part-of-speech tagged texts were used for the following linguistic features: agentless passive, attributive adjective, adverb, *be* as main verb, *by* passive, conjunct, contraction, demonstrative pronoun, direct WH-question, discourse particle, emphatic, hedge, independent clause coordination, noun, other adverbial subordinator, past participial postnominal (reduced relative) clause, past tense, perfect aspect declarative, phrasal coordination, possibility modal, predicative adjective, predictive modal, preposition, present tense, pro-verb *do*, split auxiliary, split infinitive, stranded preposition, synthetic negation, *that* clauses controlled by a verb, *that* clauses controlled by an adjective, *that* relatives on object position, *that* relatives on subject position, *to*-clause, WH-clause, WH-relative on object position, WH-relative on subject position, and WH-relatives with fronted preposition.

The following rules were applied to the programming scripts to achieve a high accuracy rate for the frequency counts. First, scripts were set to the case insensitive setting. However, when the programming script is set to the case insensitive setting, frequency counts of, for example, *us*, *Us*, *uS*, and *US* are automatically counted, even when the target form is the personal pronoun *us*, and not, for example, the proper noun *US*. To solve this unique problem, the frequencies of each *us*, *Us*, *uS*, and *US* were counted separately, and by reading the context of each word, its part-of-speech was checked manually. Second, as contractions were not expanded during data cleaning, they were searched for in the contracted form. Third, plural forms were included in all

programming scripts. Additionally, a word for which it was difficult to be consistent in determining their grammatical functions (e.g., It was difficult to determine whether *so* was used as a conjunction or an adverb) was not included in the analysis.

Before marking up the data with part-of-speech information, selecting the optimal part-of-speech tagger for the present study was difficult given the broad range of choices with dissimilar tagsets. In general, however, POS taggers can be divided into three types: rule-based, probability-based, and a combination of rule- and probability-based taggers (Hunston, 2002). Most of these have an accuracy rate that is greater than 95% with tagsets varying from 50 to more than 250 tags indicating part-of-speech classes (Garside, Leech, & McEnery, 1997). Although the tagger can be chosen by the precision rates of identifying part-of-speech information, they all have fairly high accuracy rates and are suitable for practical use. The number of tagsets can also be the basis for choosing an optimal tagger. On the one hand, it is plausible to use a minutely subdivided tagset for tagging the production of advanced learners or native English speakers, which is relatively complicated, and a roughly subdivided tagset for tagging the performance of novice and intermediate learners, which is relatively simple. However, as each POS tagger is based on a different approach to classifying part-of-speech information, it is unreasonable to choose one only by the number of tagsets.

What, then, are the key criteria that should be employed when selecting an appropriate POS tagger? As annotation of learner language is involved in the present study, numerous lexical errors and unnatural sentence structures, which might stop the POS tagger from maintaining acceptable precision rates, must be considered. It would be

ideal to automatically convert erroneous learner language into corrected target language before inserting the part-of-speech information, but this process cannot be implemented appropriately until a computer-aided error detection and correction system is fully developed. In this study, therefore, the accuracy rate of the POS tagger is maintained by manually detecting and correcting the inaccuracies of part-of-speech tagging rather than comparing the performance of various POS taggers to choose the optimal one.

A tagging program, the TreeTagger (Schmid, 1994), was employed. It is a probabilistic POS tagger that attains a 96.36% accuracy rate in predicting part-of-speech information (Schmid, 1994). The 48 tags that this tagger uses are mostly derived from the Penn-Treebank tagsets (<https://www.cis.upenn.edu/~treebank/>). However, some tags are different; the verb category, for example, is divided into three subdivisions, *be-*, *have-*, and other verbs (Appendix B). Before starting the part-of-speech tagging, a text editor was used to pre-process the data into a one-word per line format and to separate punctuation from words. A batch file was also prepared to process multiple files because TreeTagger does not execute more than one file in its default setting. Sample output of the part-of-speech tagged data is shown in Table 12. In this table, the first column shows the original text, the second column indicates the POS tag, and the third column displays the lemma form. The first line in this sample, *I* is annotated with a POS tag, PP, which stands for a personal pronoun.

Table 12. *Sample POS Tagged Data*

Word	POS	Lemma
I	PP	I
'm	VBP	be
Doing	VVG	do
Very	RB	very
Well	RB	well
Today	NN	today

In order to maintain the consistency of annotating 1,263 text files and achieve precise frequency counts, the performance of the POS tagger was examined by creating sample files. First, eight sample files that each contained 150-word passages were randomly chosen from levels 3 to 9 as well as from the data produced by the native speakers of English, totaling 1,200 words. After these eight sample files were annotated by the POS tagger, erroneously tagged items were manually counted to calculate the accuracy rate of the POS tagger. However, for some of the part-of-speech tagged items, it was difficult to judge whether they were correctly tagged or not for two reasons. The POS tagger predicts the part-of-speech information from the immediate context, that is, the previous word or tag, and the following word or tag (Schmid, 1994), but this prediction can be erroneous because of the deficiency of the tagger itself or learner language errors that make it difficult for the POS tagger to predict patterns accurately. Additionally, there are cases in which more than one part-of-speech can be assigned to a word, making it difficult to decide if the POS tagging result is correct or not (Santorini, 1991). Before presenting the accuracy rate of the tagger for each English learner proficiency level and the native speakers of English, the POS tagged L2 data examples and problematic cases were indicated as follows: (a) grammatically deviant L2 utterances,

(b) grammatically incorrect L2 utterances tagged by a grammatically correct tag, and (c) problematic cases in assigning the POS tag.

### Problems in the Part-of-Speech Tagging

First, I describe grammatically deviant L2 utterances. Table 13 shows the examples that were tagged with a grammatically incorrect and correct POS tag. These examples are from the NICT JLE corpus. As shown in Table 13, the POS tagger recognized the word *visit* as a NN (singular noun) because the learner used the wrong verb form, when the present participle should have been used. However, if the learner had used the correct verb form *shopping*, the POS tagger would have assigned the correct POS tag, VVG (present participle). Considering these two examples, the grammaticality of the POS tag in the context was not considered in this evaluation of the tagger results, as the tags were caused by non-native speaker errors and not by any deficiency in the POS tagger.

Table 13. *Learner Data Tagged with a Grammatically Incorrect POS Tag*

Words	POS	Words	POS
I	PP	And	NP
Have	VHP	I	PP
Been	VBN	Have	VHP
Visit	NN	Been	VBN
To	TO	shopping	VVG
Las	NP	In	IN
Vegas	NP	America	NP

In other cases, a grammatically incorrect L2 utterance is tagged by a grammatically correct tag. Unlike the aforementioned example, the POS tagger

sometimes assigns a grammatically correct tag to incorrect L2 production. Table 14 is an example sentence that was extracted from the NICT JLE corpus. It is tagged by a grammatically correct POS tag (“I have been to visit China and Turkey and England and French and Italy.”). The word *French* cannot be categorized as a NP (proper noun) because it is not in the proper noun form *France*. However, the POS tagger might have assigned a grammatically correct tag by considering the context in which proper nouns are juxtaposed with coordinating conjunctions. In general, it is difficult to determine what information the POS tagger used to decide the part-of-speech tag, but when grammatically incorrect words were automatically assigned grammatically correct tags, these were not counted as POS tagging errors.

Table 14. *Learner Data Tagged by a Grammatically Correct POS Tag*

Words	POS
I	PP
Have	VHP
Been	VCN
To	TO
Visit	VV
China	NP
And	CC
Turkey	NP
And	CC
England	NP
And	CC
French	NP
And	CC
Italy	NP

Finally, problematic cases in assigning the POS tag also occurred in the data.

Apart from cases involving learner errors, there are cases in which it is difficult to decide if the POS tags are assigned correctly or not. These problematic cases were checked by

referring to examples and protocols provided in the Part-of-Speech Tagging Guidelines (Santorini, 1991). The confusing tags dealt with in the present study are summarized in Table 15. The following examples show how “nouns (NN) and adjectives (JJ)” and “adjectives (ADJ) and gerunds / past participles (VVN)” are problematic cases in assigning POS tag.

Table 15. *Examples of Problematic Cases in Assigning POS Tag*

---

1. Nouns (NN) or adjectives (JJ)	<p>“Nouns that are used as modifiers, whether in isolation or in sequences, should be tagged as nouns (NN, NNS) rather than as adjectives (JJ)” (Santorini, 1991, p. 12).</p> <ul style="list-style-type: none"> <li>• computer/NN systems (4_0313.txt)</li> <li>• station/NN man (6_0788.txt)</li> <li>• animation/NN industry (NS_0007.txt)</li> <li>• *ski/JJ area (5_570.txt)</li> <li>• *ski/JJ instructor (5_570.txt)</li> <li>• *ski/JJ gerund (5_570.txt)</li> </ul>
2. Adjectives (JJ) or gerunds / past participles (VVN)	<p>The following words are considered to be adjectives (JJ): “If it is gradable—that is, if it can be preceded by a degree adverb like <i>very</i>, or if it allows the formation of a comparative” (Santorini, 1991, p. 16).</p> <ul style="list-style-type: none"> <li>• He was very surprised /JJ. (Santorini, 1991, p. 16)</li> <li>• In this week, I was very tired/JJ. (3_0155.txt)</li> <li>• I'm very *surprised/VVN. (4_0185.txt)</li> </ul>

---

*Note.* The problematic cases were shown by adding an asterisk in the examples. NS = Native speaker of English.

Other than these examples, the word *so*, which appears frequently in this spoken database, was difficult to classify in terms of its part-of-speech. It was often used as an interjection in incomplete sentences without a clear context, so it was not possible to determine whether or not it was used as a conjunction or as an adverb. Therefore, *so* was not included when calculating the accuracy rate of the POS tagging. This exclusion might

slightly skew the estimates of the accuracy rate, but as the word *so* was not targeted in this study, it cannot cause significant problems.

### **Accuracy Rate of the Part-of-Speech Tagging**

Considering the aforementioned problematic cases, POS tagging errors were counted manually and translated into the coverage of one error in every X words. Table 16 shows examples of POS tagging errors and the POS tags that are considered to be correct in randomly chosen eight sample files that each contained 150-word passages. It also shows the percentage of error rate in each oral proficiency levels by the number of times that one error appears in every X words. The sample files used for calculating the accuracy rate of the POS tagging cannot represent the whole spoken corpus data, but the following results indicate a fairly high overall accuracy in part-of-speech annotation.

The accuracy rate of the POS tagging was checked in order to maintain consistency in the annotation and to attain precise frequency counts. Inaccurately tagged items are supposed to be manually edited, but TreeTagger achieved more than a 95% accuracy rate for almost all levels as shown in Table 16 (level 3: 98.00%, level 4: 97.33%, level 5: 96.67%, level 6: 97.33%, level 7: 99.33%, level 8: 97.33%, level 9: 94.67%, NS: 98.00%).

Consequently, TreeTagger was considered an appropriate POS tagger for annotating learner language in this study. TreeTagger was unable to identify some proper nouns correctly. Proper nouns related to personal information (e.g., the name of a person, place, or company) were all replaced with XXX01 to XXX25 tags to protect the test-takers'

Table 16. *The POS Tagging Error Coverage of One Error in Every X Words*

Level	Inaccurate POS tag	Correct POS tag	1 error in every X words
3	1. I enjoyed shopping_NN in America. 2. good-bye_JJ 3. I cooking_NN rice and miso soup,	VVG NN VVP	50 (98.00%)
4	1. Eldest_JJ is seven 2. such kind of comment in Japanese_JJ, 3. she's very kind_NN, 4. Maybe because_RB,	JJS NP JJ IN	37.5 (97.33%)
5	1. I use to go skiing_NN 2. ski_JJ area 3. ski_JJ instructor 4. ski_JJ gerende 5. No waiting_VVG time	VVG NN NN NN NN	30 (96.67%)
6	1. I think that_WDT 's dangerous 2. My_NP God 3. I mean that_WDT was OK. 4. It's really really crazy crowded_VVN	DT PPS DT JJ	37.5 (97.33%)
7	1. I was tired_VVN because of football.	JJ	150 (99.33%)
8	1. He actually is tasting wine, sorry_RB. 2. but someone who actually kind_RB of 3. The restaurant that_IN I really like is 4. I just order_VV kind of mainly food,	JJ NN WDT VVP	37.5 (97.33%)
9	1. or chase_VV your friends and *have_VHP pizzas 2. kids are *kind_JJ of it's kind_NN of 3. two girls planned_VVN 4. to go camping_NN in the mountains 5. And ended_VVN up in the hotel 6. *drinking_NN coffee and looking_VVG at 7. And then *tied_VVN strings on them and put_VV 8. And then tied_VVN strings on them and *put_VV	VV NN VVD VVG VVD VVG VVD VVD	18.75 (94.67%)
NS	1. other center-sponsored activities as_IN well 2. you sort_RB of get a full view of 3. also the one_CD that has cartoons	RB NN NN	50 (98.00%)

*Note.* The inaccurate POS tags are shown by adding an asterisk in the examples. NS = Native speaker of English.

privacy. TreeTagger identified these concealed types of proper nouns as adjectives in 3,153 cases, and correctly identified them as proper nouns in 5,214 cases. In addition to this, TreeTagger correctly identified *that* as a subordinating conjunction when it introduces complements of nouns (i.e., the fact that they are not working on homework),

but the output contained an unnecessary slash mark (/that). These POS tag errors and output deficiencies were corrected manually, but they were not checked by a second person.

### **Frequency Counting**

After part-of-speech information was added, the frequency of linguistic features was automatically counted. The total frequency of each linguistic feature was double-checked using the free corpus software AntConc version 3.2.1 (2007). The frequency information of 58 linguistic features used by 1,263 test-takers across eight different oral proficiency groups was quantitatively analyzed in this study. As a result, a total of 936,320 linguistic samples were automatically extracted from the NICT JLE corpus, so that it is impossible to check manually if all of samples were correctly extracted.

The automatic part-of-speech tagging can differentiate the same lexical items which are used as different parts-of-speech. In the case of *until*, for example, if the word precedes noun phrases it can be recognized as a preposition, but if the word precedes a subject and verb structure it can be recognized as a conjunction. On the other hand, in the case of modal verb *may*, for example, it can be used for asking permission or expressing possibility, but the automatic part-of-speech tagger cannot distinguish this functional difference. However, in the case of modal verb *may*, the Biber's (1988) original list does not distinguish the functional difference. In the present study, Tree Tagger achieved a 97% accuracy rate on average (level 3: 98.00%, level 4: 97.33%, level 5: 96.67%, level 6:

97.33%, level 7: 99.33%, level 8: 97.33%, level 9: 94.67%, NS: 98.00%), but the flaws of automatic part-of-speech tagging and frequency counting were revised by hand as much as possible. However, it is impossible to check if the 936,320 linguistic samples which were automatically extracted from the NICT JLE corpus were based on the correctly POS tagged data. If there was a second person to check the items I checked manually, the results of his or her check with my own can be used to get some sort of estimate of my own error rate. Such a check would be useful in future studies.

### **Statistical Analyses**

Biber (1988) aimed to distinguish linguistic variation between written and spoken processing modes multi-dimensionally, so he selected a list of linguistic features carefully based on previous research on spoken and written language use. Then, he used exploratory factor analysis to identify common factors (i.e., dimensions) which could statistically explain the variation across different processing modes. I use the same linguistic features which are used in Biber (1988) to interpret the variation across different oral proficiency groups, but because my research goal is different, I do not use Biber's dimensions. Instead, I used different statistical analyses, namely box-and-whisker plot analysis and correspondence analysis.

### **Box-and-Whisker Plot Analysis**

The box-and-whisker plot analysis simply examines medians and percentile scores. In this study, the box-and-whisker plots can show how the linguistic features are

distributed across the oral proficiency groups. The bottom, middle and top lines of each box indicate the 25th, the 50th, and the 75th percentiles in the distribution of frequencies across the learners within the oral proficiency level. The upper whisker is at a point 1.5 times the upper quartile range above the box and the lower whisker is at a point 1.5 times the lower quartile range below the box. The size of the boxes, the length of the whiskers, and the number of outliers allow a more detailed examination of each grammatical feature.

### **Correspondence Analysis**

Correspondence analysis is useful in analyzing large amounts of data. The data to be used in such an analysis are best first presented in a data table of rows and columns. This is called a contingency table. In the case of this study there are 58 rows and eight columns in the contingency table. In a much simpler table of say six rows and three columns it would be relatively easy to interpret the eighteen data points (6 times 3) by just looking at the table and noting trends and patterns. With 58 rows and eight columns this is not easy, so it is necessary to have some way of analyzing the data and presenting it in a graphical display. That is to say, the purpose of using correspondence analysis is, according to Husson et al (2011), “to identify the primary characteristics” of a contingency table. In other words, it aims to simplify the data to facilitate interpretation.

The data in the present study use non-negative data on the same scale (SST frequency of occurrence) and this meets the requirements of correspondence analysis. Some linguistic features in Biber (1988) were not frequently used even by native

speakers of English in the present study, and this low frequency phenomenon might present a serious problem for corpus-based studies. Low frequency items are not well represented in the corpus unless the researcher obtains a huge amount of data, and these low frequency items can skew the results of the statistical analysis. Fortunately in this study, there were multiple occurrences of each grammatical feature at each proficiency level, with only one feature not occurring at all at one proficiency level (WH relative with fronted preposition at Level 3), and many of the occurrences of the features at each level being in at least double figures. However, data were in fact quite rare for a few features.

The first step in correspondence analysis is to establish if there is a significant dependency between the rows and columns (Bendixen, 2003). Typically Pearson's chi-square test for independence is used to see if the null hypothesis can be rejected, namely there is no relationship between rows and columns. In the present study, I expect to see an interdependency signaling that certain grammatical features distinguish particular oral proficiency levels. The second step in correspondence analysis is to check the number of dimensions (also called factors) needed to describe the points and to see how much of the chi-square value they account for. In the present study a strong oral proficiency level is expected to be a factor. The greater the amount of data accounted for by the factors, the better the original data will be represented in the plot. The next step is to produce the graphical representation of the data in the contingency table. The statistical program does this using chi-square distances by calculating and plotting row scores and then by calculating and plotting column scores.

Correspondence analysis has two advantages over other multivariate statistical methods. The first advantage lies in its calculation, which is similar to other multivariate statistical analyses (e.g., Hayashi's Quantification Method Type III, dual scaling method). Its calculation is simple because the mathematical solution can be gained by one eigenvalue calculation and singular value decomposition. What is more, there is no option in the calculation process; thus, its reproducibility is high compared with factor analysis and principal component analysis (Kobayashi, 2010). The calculation for both factor analysis and principle component analysis is based on a correlation matrix that is derived from a frequency table, but correspondence analysis directly uses a frequency table to calculate quantities for categories and samples to examine the degree of dependency (Nakamura, 2005). Those quantities are then used to analyze the characteristic features of the categories and samples. As the quantities are calculated, a correlation coefficient between the quantities of categories and samples are supposed to be maximized. In other words, correspondence analysis maximizes the canonical correlation of rows and columns in the frequency table and rearranges the matrix of rows and columns to produce the highest values around the diagonal (Kobayashi, 2010; Tabata, 2002). Thus, rearranged categories and samples in close proximity are considered to be qualitatively similar. The second and major advantage of correspondence analysis is that it shows similarities and dissimilarities among variables (the linguistic features) and cases (the oral proficiency levels) in a joint plot. Having all the items in one display reduces the complexity of the data. Once this overall picture has been established, particular items can then be chosen for detailed investigation (Baayen, 2008).

What is more, correspondence analysis is considered to be a useful statistical approach to describe text variation quantitatively in previous studies. Nakamura (2002) concluded that three statistical methods are most likely to accomplish this purpose: (a) factor analysis, (b) principal component analysis, and (c) quantification in a contingency table (i.e., correspondence analysis). Among these three statistical methods, Nakamura (2002) recommended correspondence analysis because it is powerful but not complicated when processing a large number of data points. Mizumoto (2009) compared the results elicited from principal component analysis, correspondence analysis, and principal component analysis with a transposed matrix. He concluded that correspondence analysis, which involves making a representation of the data from a contingency table, is the most suitable method for investigating the similarity and dissimilarity of variables because principal component analysis is used for continuous data while correspondence analysis is used with non-continuous data. Tabata (2004) arrived at the same conclusion.

### **Procedures**

The following procedures were used in this study:

1. Fifty-eight linguistic features out of 67 listed in Biber (1988) were chosen to identify the linguistic features that can specify the characteristics of different oral proficiency levels.
2. The part-of-speech (POS) information was marked automatically with TreeTagger (Schmid, 1994). The 48 tags this tagger uses are mostly derived from the Penn-Treebank tagsets (Appendix B). As TreeTagger was not able to correctly detect

proper nouns that were replaced with XXX01 to XXX25 tags to protect test-takers' privacy, they were corrected before the target linguistic features were counted.

3. The raw frequencies of features were automatically counted using a computer programming language, Perl, which has been often employed for linguistic analysis (Hammond, 2003). The patterns for the programming script were mostly adopted from Biber (1988) (Appendix C), but as the original one includes some defects as pointed out in de Mönnink, Brom, and Oostdijk (2003), minor changes that can improve the precision and recall rate (measure of the accuracy of information extraction) were added. These changes are described in Appendix C. As an example, when the raw frequency of direct WH-questions are counted, WHP and V part-of-speech tags were added to those used by Biber (1988). As suggested in Murakami (2009), the original version fails to retrieve some question sentences such as "*Who went there?*" However, these additions can help to retrieve *wh* questions better. The programming scripts were modified based on the scripts developed by Murakami (2009).
4. The raw frequencies of features were double-checked using the same search formula through the free corpus software AntConc version 3.2.1 (2007). Additionally, concordance lines were checked to confirm if the targeted linguistic features were extracted properly and non-targeted linguistic features were not included. Afterwards, some search formulas were modified to fit the characteristics of learner language to increase the precision and recall rate in frequency profiling. Because of learner errors of not using the correct verb form, past participles were recognized as past tenses in some cases (e.g. I have wrote). Therefore, past tense "VBD" tag was added to extract

the perfect aspect from the learner data. Then, a frequency matrix of eight oral proficiency levels x 58 linguistic features was created.

5. Some of the grammatical features are focused on by looking at the descriptive statistics displayed in box-and-whisker plots. The notches on the box-and-whisker plots (i.e., confidence intervals) are used to check where the significant differences are.
6. Correspondence analysis was performed using the statistical software R version 2.10.1 (2009). The R scripts used for the present analysis are provided in Appendix D.

## **CHAPTER 4**

### **RESULTS**

In the present study, I examined grammatical features that can distinguish particular oral proficiency levels. The first approach involves examining box-and-whisker plots of medians and percentile scores using data that preserve each individual's use of grammatical features normalized for frequencies per 100 words. The second approach is called correspondence analysis that uses a complex statistical procedure to make a graphical plot to understand the characteristics of a contingency table. The data used in the procedure consists of a large number of transcribed spoken texts joined together to make a sub-corpus for each of seven oral proficiency levels. These two statistical analyses differ in that the box-and-whisker plots allow a detailed examination of each grammatical feature and correspondence analysis provides an overall picture of the grammatical features. These two approaches thus nicely complement each other, with box-and-whisker plots being a useful follow up to correspondence analysis.

#### **Results of Descriptive Statistics**

This study is an exploratory study, so I begin by presenting descriptive statistics (e.g., tables of means, standard deviations, medians, and box-and-whisker plots) before moving on to more sophisticated analyses such as Correspondence analysis. Table 17, Table 18, Table 19 and Table 20 provide the means, standard deviations, and medians for both normalized data and raw data. The content of descriptive statistics is linked to the

box-and-whisker plots as they all complement each other, so I draw out a few highlights to illustrate the patterns that will be focused on in the following sections.

First, *noun* (E15) is the most frequently used linguistic feature in this study, so it has high medians across the oral proficiency levels. It ranges from 25.42 (L3) occurrences per 100 words to 17.48 (NS) occurrences. There is a falling pattern of medians through the oral proficiency levels, and this pattern can be a sign of increasing proficiency. In contrast, *emphatic* (J42) has low medians across the oral proficiency levels, which range from 0.27 (L3) occurrences per 100 words to 2.40 (NS). There is an increase pattern of medians through the levels, and this pattern can be a sign of increasing proficiency. Additionally, *pronoun it* (C9) has a slightly higher frequency than *emphatic*, but there are still low medians across the oral proficiency levels. It ranges from 0.75 (L3) occurrences per 100 words to 2.66 (NS). There is also an increase pattern of medians through the levels as in *emphatic*. The medians of *infinitive* (H23) and *possibility modal* (K44) range between a narrow bound. *Infinitive* ranges from the lowest 1.01 (L3) occurrences per 100 words to the highest 1.88 (L9) and *possibility modal* ranges from the lowest 0.24 (L3) occurrences per 100 words to the highest 0.82 (L9). Level 3 learners and native speakers of English have lower medians of these two features compared with the other oral proficiency levels.

Table 17. Descriptive Statistics for Level 3, 4, 5, and 6 (Normalized Frequency per 100 Words)

		Level 3			Level 4			Level 5			Level 6		
		M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn
A1	past.tense	2.06	1.18	1.90	2.59	1.11	2.48	3.17	1.12	2.97	3.55	1.29	3.38
A2	perfect.aspect	0.10	0.20	0.00	0.13	0.19	0.00	0.15	0.16	0.12	0.19	0.21	0.14
A3	present_tense_final	14.63	2.44	14.65	13.96	2.05	13.83	12.95	1.92	13.07	13.06	1.87	13.08
B4	place_adv	0.36	0.37	0.27	0.35	0.30	0.29	0.37	0.27	0.32	0.34	0.24	0.31
B5	time.adverb	0.66	0.51	0.55	0.69	0.44	0.62	0.73	0.43	0.67	0.69	0.38	0.63
C6	first.person.pronoun	8.83	1.87	8.90	8.58	1.60	8.51	8.52	1.61	8.48	8.48	1.56	8.35
C7	scndpsn_pro	1.51	0.74	1.50	1.32	0.58	1.27	1.25	0.59	1.17	1.43	0.86	1.22
C8	thrdpsn_pro	2.35	1.46	2.22	2.37	1.25	2.25	2.48	1.21	2.40	2.45	1.14	2.36
C9	it.pronoun	0.93	0.79	0.75	1.22	0.75	1.08	1.47	0.80	1.42	1.63	0.72	1.48
C10	demonstrative.pronoun	0.29	0.36	0.22	0.38	0.35	0.30	0.44	0.33	0.36	0.57	0.35	0.51
C11	indefinite.pronoun	0.19	0.29	0.00	0.23	0.27	0.16	0.29	0.27	0.24	0.44	0.34	0.38
C12	proverb.do	0.23	0.31	0.18	0.25	0.25	0.18	0.25	0.24	0.20	0.29	0.26	0.26
D13	wh_question	0.30	0.34	0.22	0.25	0.23	0.20	0.19	0.19	0.14	0.18	0.20	0.12
E14	nominal	0.67	0.54	0.57	0.88	0.51	0.80	0.95	0.50	0.88	0.95	0.55	0.90
E15	noun	26.14	3.99	25.42	22.29	2.65	22.11	20.09	2.02	20.00	18.58	1.92	18.45
F16	agentless.passive	0.18	0.23	0.14	0.20	0.21	0.17	0.29	0.25	0.25	0.34	0.21	0.32
F17	by.passive	0.01	0.05	0.00	0.01	0.05	0.00	0.02	0.05	0.00	0.03	0.07	0.00
G18	be_main_verb	1.42	0.70	1.36	1.35	0.58	1.29	1.29	0.51	1.27	1.28	0.46	1.20
G19	exist_there	0.46	0.48	0.33	0.50	0.41	0.43	0.48	0.40	0.40	0.42	0.35	0.38
H20	that_clause_by_verb	0.02	0.07	0.00	0.04	0.08	0.00	0.06	0.10	0.00	0.09	0.12	0.08
H21	that_clause_by_adjective	0.00	0.02	0.00	0.00	0.02	0.00	0.01	0.03	0.00	0.01	0.04	0.00
H22	WH.clause	0.03	0.09	0.00	0.05	0.11	0.00	0.08	0.11	0.00	0.09	0.10	0.09
H23	infinitive	1.08	0.70	1.01	1.64	0.74	1.61	1.87	0.72	1.74	1.79	0.55	1.74

(Table 17 continues).

(Table 17 continued).

		Level 3			Level 4			Level 5			Level 6		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
H24	past.participial.postnominal.clause	0.02	0.06	0.00	0.03	0.08	0.00	0.03	0.06	0.00	0.04	0.06	0.00
H25	that_relative_subject	0.01	0.05	0.00	0.01	0.05	0.00	0.01	0.04	0.00	0.01	0.05	0.00
H26	that_relative_object	0.01	0.07	0.00	0.03	0.08	0.00	0.03	0.06	0.00	0.03	0.07	0.00
H27	wh_relative_subject	0.01	0.07	0.00	0.05	0.11	0.00	0.09	0.13	0.00	0.13	0.18	0.08
H28	WH.relative.in.object.position	0.01	0.04	0.00	0.01	0.05	0.00	0.02	0.06	0.00	0.03	0.07	0.00
H29	wh_relative_front_prep	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.01	0.03	0.00
H30	causative.adverbial.subordinator	0.25	0.35	0.19	0.45	0.35	0.37	0.56	0.34	0.51	0.60	0.38	0.58
H31	conce_adv_sub	0.01	0.05	0.00	0.01	0.03	0.00	0.01	0.04	0.00	0.01	0.04	0.00
H32	coconditional.adverbial.subordinator	0.05	0.13	0.00	0.12	0.17	0.00	0.19	0.19	0.15	0.23	0.20	0.20
H33	other_adverbial_subordinator	0.05	0.16	0.00	0.07	0.15	0.00	0.11	0.17	0.00	0.20	0.25	0.11
I34	prepo	5.76	1.76	5.90	6.68	1.38	6.76	6.98	1.28	7.03	6.78	1.20	6.76
I35	attributive.adjective	5.34	1.53	5.22	5.47	1.25	5.47	5.42	1.12	5.28	5.16	0.98	5.21
I36	predicative_adjective	0.79	0.54	0.75	0.74	0.44	0.67	0.73	0.39	0.71	0.72	0.34	0.66
I37	adverb	3.44	1.67	3.23	4.41	1.67	4.33	5.08	1.55	4.97	5.36	1.77	5.31
J38	conjunct	0.07	0.14	0.00	0.12	0.17	0.00	0.14	0.19	0.10	0.13	0.22	0.00
J39	downtoner	0.13	0.21	0.00	0.18	0.22	0.14	0.16	0.19	0.12	0.14	0.13	0.11
J40	hedge	0.23	0.36	0.00	0.34	0.39	0.21	0.42	0.41	0.31	0.46	0.42	0.36
J41	amp	1.11	0.90	0.91	1.11	0.66	1.04	1.17	0.80	1.05	1.09	0.61	1.06
J42	emphatic	0.46	0.54	0.27	0.60	0.45	0.51	0.84	0.51	0.76	1.08	0.58	0.96
J43	discourse_particle	0.11	0.19	0.00	0.17	0.23	0.12	0.18	0.20	0.12	0.18	0.20	0.12
K44	possibility.modal	0.33	0.36	0.24	0.53	0.42	0.45	0.76	0.41	0.70	0.83	0.38	0.80
K45	necessity.modal	0.06	0.14	0.00	0.07	0.14	0.00	0.08	0.14	0.00	0.12	0.16	0.09
K46	predictive.modal	0.36	0.37	0.27	0.49	0.40	0.43	0.48	0.36	0.41	0.43	0.30	0.40

(Table 17 continues).

(Table 17 continued).

		Level 3			Level 4			Level 5			Level 6		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
L47	public.verb	0.26	0.30	0.20	0.29	0.30	0.20	0.36	0.29	0.32	0.46	0.35	0.41
L48	private.verb	0.98	0.68	0.82	1.31	0.67	1.26	1.51	0.70	1.44	2.00	1.03	1.69
L49	suasive.verb	0.12	0.19	0.00	0.24	0.27	0.17	0.25	0.21	0.21	0.28	0.23	0.24
L50	seem.appear	0.01	0.07	0.00	0.03	0.08	0.00	0.04	0.10	0.00	0.07	0.13	0.00
M51	contraction	2.01	1.02	1.79	2.21	0.89	2.16	2.40	0.99	2.27	2.93	0.95	2.79
M52	stranded.preposition	0.19	0.25	0.00	0.14	0.16	0.13	0.14	0.15	0.12	0.16	0.15	0.13
M53	split_infinitive	0.00	0.05	0.00	0.00	0.03	0.00	0.00	0.02	0.00	0.01	0.03	0.00
M54	split_auxiliary	0.07	0.15	0.00	0.09	0.12	0.00	0.13	0.14	0.11	0.17	0.15	0.12
N55	phrasal_coordination	0.88	0.57	0.81	0.58	0.42	0.49	0.46	0.31	0.41	0.36	0.26	0.33
N56	ind_clause_coordination	2.05	1.31	1.65	2.53	1.19	2.42	2.44	1.02	2.29	2.41	1.04	2.39
O57	synt_negation	0.17	0.23	0.00	0.12	0.16	0.00	0.10	0.14	0.00	0.10	0.12	0.08
O58	analytic.negation	0.96	0.71	0.85	1.08	0.54	1.04	1.17	0.56	1.11	1.58	0.57	1.46

Table 18. Descriptive Statistics for Level 7, 8, 9, and Native Speakers of English (Normalized Frequency per 100 Words)

		Level 7			Level 8			Level 9			NS		
		M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn
A1	past.tense	4.28	1.16	4.38	4.10	1.42	3.95	4.39	1.34	4.40	3.42	0.94	3.32
A2	perfect.aspect	0.20	0.17	0.15	0.20	0.15	0.19	0.24	0.19	0.19	0.50	0.18	0.50
A3	present_tense_final	13.03	1.84	13.16	12.99	1.91	13.02	13.22	1.71	13.13	13.29	1.62	13.11
B4	place_adv	0.35	0.23	0.35	0.31	0.20	0.30	0.39	0.24	0.33	0.52	0.17	0.55
B5	time.adverb	0.66	0.38	0.55	0.57	0.30	0.58	0.55	0.24	0.50	0.41	0.17	0.43
C6	first.person.pronoun	8.84	1.41	8.94	8.35	1.46	8.43	8.26	1.58	8.31	7.09	0.99	6.80
C7	scndpsn_pro	1.48	0.82	1.29	1.64	1.09	1.41	1.61	0.68	1.48	1.49	0.48	1.48
C8	thrdpsn_pro	2.25	1.02	2.17	2.31	0.94	2.31	2.67	0.92	2.61	2.35	0.51	2.34
C9	it.pronoun	2.00	0.71	2.03	2.00	0.77	1.95	2.13	0.70	2.01	2.66	0.56	2.66
C10	demonstrative.pronoun	0.55	0.29	0.53	0.59	0.30	0.58	0.60	0.28	0.56	0.60	0.12	0.57
C11	indefinite.pronoun	0.43	0.32	0.37	0.33	0.26	0.24	0.48	0.26	0.47	0.54	0.23	0.48
C12	proverb.do	0.43	0.29	0.42	0.40	0.29	0.34	0.52	0.29	0.54	0.54	0.16	0.49
D13	wh_question	0.12	0.14	0.09	0.15	0.19	0.10	0.14	0.15	0.09	0.19	0.13	0.15
E14	nominal	0.80	0.44	0.73	0.80	0.41	0.76	0.72	0.43	0.58	0.84	0.29	0.81
E15	noun	17.97	1.48	17.93	17.66	1.80	17.47	17.08	1.38	16.98	17.31	1.38	17.48
F16	agentless.passive	0.39	0.26	0.35	0.39	0.24	0.38	0.36	0.24	0.34	0.32	0.14	0.32
F17	by.passive	0.03	0.06	0.00	0.03	0.04	0.00	0.02	0.05	0.00	0.02	0.03	0.01
G18	be_main_verb	1.26	0.46	1.27	1.24	0.47	1.11	1.33	0.38	1.28	1.29	0.28	1.34
G19	exist_there	0.40	0.35	0.27	0.41	0.31	0.34	0.44	0.33	0.33	0.54	0.17	0.52
H20	that_clause_by_verb	0.15	0.15	0.10	0.20	0.23	0.11	0.18	0.16	0.13	0.14	0.06	0.13
H21	that_clause_by_adjective	0.03	0.06	0.00	0.03	0.05	0.00	0.03	0.05	0.00	0.03	0.03	0.03
H22	WH.clause	0.15	0.14	0.12	0.13	0.12	0.09	0.15	0.11	0.14	0.15	0.09	0.13
H23	infinitive	1.88	0.64	1.80	1.81	0.60	1.66	2.04	0.56	1.88	1.50	0.32	1.50
H24	past.participial.postnominal.clause	0.02	0.05	0.00	0.03	0.05	0.00	0.03	0.05	0.00	0.03	0.02	0.03

(Table 18 continues).

(Table 18 continued).

		Level 7			Level 8			Level 9			NS		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
H25	that_relative_subject	0.03	0.06	0.00	0.03	0.04	0.00	0.07	0.09	0.00	0.13	0.07	0.14
H26	that_relative_object	0.08	0.12	0.00	0.10	0.14	0.00	0.14	0.12	0.13	0.20	0.11	0.19
H27	wh_relative_subject	0.12	0.15	0.08	0.14	0.17	0.10	0.19	0.18	0.17	0.14	0.06	0.14
H28	WH.relative.in.object.position	0.02	0.04	0.00	0.03	0.05	0.00	0.03	0.05	0.00	0.03	0.03	0.02
H29	wh_relative_front_prep	0.01	0.03	0.00	0.01	0.04	0.00	0.01	0.03	0.00	0.00	0.01	0.00
H30	causative.adverbial.subordinator	0.55	0.36	0.52	0.59	0.33	0.55	0.57	0.32	0.51	0.49	0.27	0.49
H31	conce_adv_sub	0.05	0.08	0.00	0.04	0.07	0.00	0.04	0.06	0.00	0.07	0.05	0.08
H32	coconditional.adverbial.subordinator	0.24	0.18	0.19	0.28	0.17	0.24	0.35	0.22	0.31	0.36	0.15	0.34
H33	other_adverbial_subordinator	0.26	0.26	0.23	0.33	0.24	0.33	0.33	0.23	0.29	0.31	0.17	0.29
I34	prepo	6.87	1.15	6.94	6.92	1.06	6.88	7.32	1.06	7.37	8.04	0.69	7.80
I35	attributive.adjective	5.12	0.99	5.04	5.04	0.92	4.90	4.76	1.01	4.81	5.22	0.58	5.20
I36	predicative_adjective	0.73	0.36	0.71	0.69	0.34	0.67	0.76	0.29	0.78	0.81	0.24	0.77
I37	adverb	5.24	1.61	5.03	5.47	1.75	5.50	4.73	1.48	4.52	5.44	0.95	5.44
J38	conjunct	0.09	0.12	0.08	0.11	0.12	0.08	0.13	0.16	0.09	0.06	0.06	0.04
J39	downtoner	0.17	0.15	0.12	0.15	0.12	0.14	0.17	0.14	0.15	0.15	0.10	0.15
J40	hedge	0.43	0.34	0.35	0.45	0.35	0.41	0.45	0.31	0.39	0.58	0.28	0.49
J41	amp	0.76	0.61	0.62	0.77	0.57	0.68	0.50	0.49	0.39	0.41	0.26	0.34
J42	emphatic	1.39	0.67	1.37	1.68	0.88	1.42	1.78	0.83	1.76	2.38	0.65	2.41
J43	discourse_particle	0.10	0.14	0.06	0.09	0.09	0.08	0.06	0.09	0.00	0.06	0.06	0.03
K44	possibility.modal	0.86	0.37	0.81	0.85	0.43	0.74	0.86	0.29	0.82	0.65	0.19	0.61
K45	necessity.modal	0.08	0.09	0.08	0.15	0.17	0.09	0.13	0.15	0.07	0.06	0.04	0.06
K46	predictive.modal	0.39	0.26	0.36	0.44	0.28	0.35	0.52	0.36	0.44	0.73	0.24	0.70
L47	public.verb	0.40	0.26	0.35	0.41	0.24	0.39	0.42	0.32	0.34	0.28	0.17	0.23
L48	private.verb	2.31	1.03	2.20	2.29	1.01	2.08	2.27	0.60	2.28	2.17	0.41	2.15

(Table 18 continues).

(Table 18 continued).

		Level 7			Level 8			Level 9			NS		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
L49	suasive.verb	0.25	0.22	0.22	0.20	0.15	0.18	0.16	0.12	0.14	0.12	0.06	0.13
L50	seem.appear	0.08	0.13	0.00	0.06	0.09	0.00	0.06	0.09	0.00	0.08	0.07	0.06
M51	contraction	3.40	1.00	3.26	3.33	0.97	3.40	3.65	1.10	3.58	4.48	0.84	4.35
M52	stranded.preposition	0.19	0.17	0.17	0.18	0.14	0.16	0.22	0.14	0.20	0.36	0.12	0.37
M53	split_infinitive	0.01	0.04	0.00	0.01	0.03	0.00	0.03	0.05	0.00	0.03	0.04	0.02
M54	split_auxiliary	0.17	0.15	0.15	0.20	0.14	0.17	0.21	0.18	0.18	0.23	0.08	0.22
N55	phrasal_coordination	0.38	0.22	0.37	0.32	0.21	0.27	0.37	0.24	0.30	0.48	0.21	0.46
N56	ind_clause_coordination	2.51	0.93	2.50	2.60	0.65	2.73	2.49	0.64	2.50	2.38	0.69	2.45
O57	synt_negation	0.07	0.09	0.00	0.06	0.10	0.00	0.06	0.08	0.00	0.08	0.05	0.07
O58	analytic.negation	1.76	0.60	1.63	1.66	0.52	1.67	1.87	0.50	1.85	1.33	0.31	1.30

Note. NS = Native speaker of English.

Table 19. Descriptive Statistics for Level 3, 4, 5, and 6 (Raw Frequency Data)

		Level 3			Level 4			Level 5			Level 6		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
A1	past.tense	9.05	5.84	8.00	16.55	7.90	16.00	27.39	11.30	25.00	35.19	13.55	34.50
A2	perfect.aspect	0.45	0.87	0.00	0.85	1.23	0.00	1.31	1.44	1.00	1.87	1.91	1.00
A3	present_tense_final	61.88	16.29	59.00	88.82	21.21	87.00	111.41	25.50	110.00	131.53	33.29	127.00
B4	place_adv	1.58	1.78	1.00	2.24	2.07	2.00	3.20	2.52	3.00	3.37	2.19	3.00
B5	time.adverb	2.85	2.33	2.00	4.44	3.05	4.00	6.32	3.82	6.00	6.96	3.94	6.00
C6	first.person.pronoun	37.85	12.47	36.00	54.54	14.10	54.00	73.28	18.59	71.50	84.59	20.51	82.00
C7	scndpsn_pro	6.41	3.41	6.00	8.40	3.99	8.00	10.89	6.02	10.00	14.91	10.93	12.00
C8	thrdpsn_pro	9.75	6.13	9.00	15.23	8.76	14.00	21.23	10.90	20.00	24.44	12.03	23.00
C9	it.pronoun	4.05	3.80	3.00	7.94	5.36	7.00	12.81	7.56	12.00	16.29	7.60	15.00
C10	demonstrative.pronoun	1.27	1.67	1.00	2.51	2.46	2.00	3.86	3.19	3.00	5.85	3.87	5.00
C11	indefinite.pronoun	0.82	1.31	0.00	1.51	1.79	1.00	2.58	2.55	2.00	4.55	3.96	3.50
C12	proverb.do	0.97	1.24	1.00	1.54	1.59	1.00	2.14	2.09	2.00	2.92	2.68	2.00
D13	wh_question	1.27	1.40	1.00	1.61	1.47	1.00	1.68	1.69	1.00	1.88	2.16	1.00
E14	nominal	2.96	2.55	2.00	5.73	3.64	5.00	8.13	4.56	8.00	9.58	5.87	9.00
E15	noun	111.05	28.85	107.00	141.09	27.79	137.50	172.72	33.19	171.00	185.74	36.55	180.50
F16	agentless.passive	0.82	1.04	1.00	1.32	1.39	1.00	2.51	2.20	2.00	3.46	2.23	3.00
F17	by.passive	0.04	0.19	0.00	0.09	0.32	0.00	0.13	0.41	0.00	0.32	0.63	0.00
G18	be_main_verb	5.98	3.11	6.00	8.56	3.83	8.00	10.94	4.51	11.00	12.84	5.25	12.00
G19	exist_there	1.94	2.03	1.00	3.19	2.70	3.00	4.00	3.37	3.00	4.12	3.14	4.00
H20	that_clause_by_verb	0.09	0.31	0.00	0.25	0.53	0.00	0.56	0.92	0.00	1.01	1.43	1.00
H21	that_clause_by_adjective	0.00	0.07	0.00	0.02	0.15	0.00	0.08	0.29	0.00	0.13	0.36	0.00
H22	WH.clause	0.16	0.45	0.00	0.36	0.73	0.00	0.72	0.97	0.00	0.93	0.96	1.00
H23	infinitive	4.73	3.39	4.00	10.56	5.43	10.00	16.10	6.74	15.00	18.06	6.43	18.00
H24	past.participial.postnominal.clause	0.08	0.30	0.00	0.19	0.48	0.00	0.25	0.55	0.00	0.41	0.59	0.00

(Table 19 continues).

(Table 19 continued).

		Level 3			Level 4			Level 5			Level 6		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
H25	that_relative_subject	0.04	0.24	0.00	0.09	0.31	0.00	0.09	0.31	0.00	0.12	0.42	0.00
H26	that_relative_object	0.05	0.26	0.00	0.18	0.52	0.00	0.25	0.57	0.00	0.32	0.71	0.00
H27	wh_relative_subject	0.06	0.30	0.00	0.34	0.78	0.00	0.81	1.19	0.00	1.31	1.91	1.00
H28	WH.relative.in.object.position	0.02	0.15	0.00	0.09	0.36	0.00	0.17	0.50	0.00	0.31	0.68	0.00
H29	wh_relative_front_prep	0.00	0.00	0.00	0.01	0.12	0.00	0.03	0.16	0.00	0.05	0.31	0.00
H30	causative.adverbial.subordinator	1.12	1.60	1.00	2.92	2.43	2.00	4.88	3.28	5.00	5.98	3.76	5.50
H31	conce_adv_sub	0.03	0.25	0.00	0.04	0.19	0.00	0.11	0.38	0.00	0.13	0.38	0.00
H32	coconditional.adverbial.subordinator	0.26	0.64	0.00	0.82	1.15	0.00	1.66	1.78	1.00	2.32	2.18	2.00
H33	other_adverbial_subordinator	0.22	0.70	0.00	0.50	1.03	0.00	0.95	1.48	0.00	1.98	2.51	1.00
I34	prepo	24.98	10.56	24.00	42.74	12.58	41.00	60.32	15.77	60.00	67.98	16.99	66.00
I35	attributive.adjective	23.39	10.15	21.50	35.26	11.60	35.00	46.88	13.60	46.00	51.99	14.63	51.00
I36	predicative_adjective	3.27	2.23	3.00	4.70	2.86	4.00	6.19	3.37	6.00	7.20	3.77	6.00
I37	adverb	14.94	9.12	13.00	28.93	14.38	27.00	44.46	18.31	42.00	54.98	24.62	51.00
J38	conjunct	0.34	0.64	0.00	0.73	1.09	0.00	1.19	1.57	1.00	1.21	2.15	0.00
J39	downtoner	0.60	0.98	0.00	1.15	1.56	1.00	1.42	1.70	1.00	1.43	1.44	1.00
J40	hedge	1.02	1.70	0.00	2.24	2.73	1.00	3.69	3.88	2.00	4.84	4.74	4.00
J41	amp	5.12	4.82	4.00	7.28	4.87	7.00	10.34	7.62	9.00	11.05	6.97	10.00
J42	emphatic	2.02	2.60	1.00	4.01	3.46	3.00	7.45	5.12	6.50	11.11	7.04	9.00
J43	discourse_particle	0.51	0.96	0.00	1.11	1.57	1.00	1.53	1.75	1.00	1.83	2.21	1.00
K44	possibility.modal	1.46	1.56	1.00	3.49	3.07	3.00	6.66	4.15	6.00	8.37	3.99	8.00
K45	necessity.modal	0.25	0.61	0.00	0.42	0.83	0.00	0.76	1.31	0.00	1.24	1.62	1.00
K46	predictive.modal	1.55	1.64	1.00	3.15	2.65	3.00	4.12	3.06	3.00	4.28	2.98	4.00
L47	public.verb	1.14	1.35	1.00	1.95	2.11	1.00	3.25	2.81	3.00	4.70	3.84	4.00
L48	private.verb	4.26	3.33	4.00	8.61	5.06	8.00	13.33	7.60	12.00	20.59	13.18	17.00

(Table 19 continues).

(Table 19 continued).

		Level 3			Level 4			Level 5			Level 6		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
L49	suasive.verb	0.50	0.81	0.00	1.58	1.74	1.00	2.17	1.98	2.00	2.74	2.15	2.00
L50	seem.appear	0.07	0.32	0.00	0.17	0.50	0.00	0.34	0.83	0.00	0.75	1.35	0.00
M51	contraction	8.50	4.81	7.50	14.26	6.76	13.00	20.81	9.53	20.00	29.34	11.02	28.50
M52	stranded.preposition	0.82	1.07	0.00	0.88	1.01	1.00	1.23	1.35	1.00	1.67	1.60	1.00
M53	split_infinitive	0.02	0.16	0.00	0.03	0.16	0.00	0.03	0.19	0.00	0.08	0.28	0.00
M54	split_auxiliary	0.32	0.70	0.00	0.58	0.82	0.00	1.18	1.27	1.00	1.68	1.54	1.00
N55	phrasal_coordination	3.75	2.48	3.50	3.59	2.44	3.00	3.82	2.37	4.00	3.44	2.20	3.00
N56	ind_clause_coordination	9.10	6.77	7.00	16.10	8.23	15.00	21.05	9.87	20.00	24.05	11.31	22.00
O57	synt_negation	0.73	0.98	0.00	0.74	1.04	0.00	0.94	1.36	0.00	0.99	1.24	1.00
O58	analytic_negation	4.01	2.96	4.00	6.94	3.93	6.00	10.11	5.24	9.00	15.81	6.37	14.00

Table 20. Descriptive Statistics for Level 7, 8, 9, and Native Speakers of English (Raw Frequency Data)

		Level 7			Level 8			Level 9			NS		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
A1	past.tense	47.00	14.40	45.00	49.02	16.50	51.50	59.23	20.87	57.00	148.10	60.58	135.00
A2	perfect.aspect	2.10	1.77	2.00	2.45	1.78	2.00	3.20	2.80	3.00	20.85	7.01	20.50
A3	present_tense_final	144.60	33.59	141.00	160.04	48.21	148.00	179.80	41.66	175.50	562.60	127.98	539.50
B4	place_adv	3.91	2.55	4.00	3.73	2.45	3.00	5.38	3.66	5.00	22.20	8.99	20.50
B5	time.adverb	7.30	4.45	6.00	7.02	4.47	7.00	7.30	2.80	7.50	17.90	10.52	16.00
C6	first.person.pronoun	97.53	21.29	100.00	100.63	21.66	97.50	111.98	29.04	108.50	300.20	74.80	280.50
C7	scndpsn_pro	16.66	10.59	14.00	21.68	21.29	17.00	22.28	11.51	19.50	63.95	26.35	60.00
C8	thrdpsn_pro	24.97	12.42	23.00	28.21	12.84	25.00	36.00	13.33	34.00	100.00	31.35	100.50
C9	it.pronoun	21.94	8.10	21.00	23.98	10.01	21.50	29.10	11.92	25.00	114.30	36.58	121.50
C10	demonstrative.pronoun	6.26	3.88	6.00	7.41	4.45	7.00	8.35	4.57	8.00	25.30	5.92	23.50
C11	indefinite.pronoun	4.95	4.08	4.00	4.02	3.29	3.00	6.65	3.66	6.00	22.85	10.60	21.50
C12	proverb.do	4.83	3.47	4.00	4.88	3.35	5.00	7.03	4.23	6.50	23.20	9.13	20.50
D13	wh_question	1.31	1.56	1.00	1.86	1.92	1.00	2.00	2.20	1.00	8.35	6.27	7.00
E14	nominal	9.16	5.78	8.00	9.93	5.26	9.00	10.00	6.56	7.50	34.95	12.45	31.50
E15	noun	198.95	38.03	198.00	216.34	51.92	206.50	231.65	43.61	241.50	729.10	132.72	702.50
F16	agentless.passive	4.25	2.77	4.00	4.95	3.56	4.50	4.95	3.64	4.00	13.55	6.03	11.50
F17	by.passive	0.35	0.77	0.00	0.30	0.50	0.00	0.28	0.64	0.00	0.95	1.15	0.50
G18	be_main_verb	13.92	5.58	13.00	14.80	5.06	14.00	17.88	6.08	17.00	55.65	19.87	50.50
G19	exist_there	4.23	3.66	3.00	4.93	3.75	4.00	5.83	4.47	4.50	23.30	9.51	22.00
H20	that_clause_by_verb	1.66	1.71	1.00	2.50	3.35	1.00	2.50	2.25	2.00	6.05	2.87	5.00
H21	that_clause_by_adjective	0.30	0.59	0.00	0.36	0.67	0.00	0.40	0.71	0.00	1.35	1.39	1.00
H22	WH.clause	1.62	1.56	1.00	1.57	1.37	1.00	2.08	1.70	2.00	6.50	4.56	5.00
H23	infinitive	20.58	7.53	19.00	22.13	8.87	22.00	27.70	9.39	25.50	62.95	14.28	66.00
H24	past.participial.postnominal.clause	0.27	0.58	0.00	0.34	0.67	0.00	0.35	0.62	0.00	1.55	1.28	1.00

(Table 20 continues).

(Table 20 continued).

	Level 7			Level 8			Level 9			NS			
	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	
H25	that_relative_subject	0.35	0.66	0.00	0.30	0.50	0.00	1.00	1.47	0.00	5.65	3.00	6.00
H26	that_relative_object	0.83	1.14	0.00	1.16	1.69	0.00	1.98	1.73	2.00	8.10	3.92	8.00
H27	wh_relative_subject	1.40	1.77	1.00	1.70	1.93	1.00	2.65	2.50	2.00	5.85	2.62	6.00
H28	WH.relative.in.object.position	0.19	0.46	0.00	0.32	0.61	0.00	0.33	0.62	0.00	1.10	1.07	1.00
H29	wh_relative_front_prep	0.06	0.30	0.00	0.09	0.44	0.00	0.15	0.43	0.00	0.15	0.37	0.00
H30	causative.adverbial.subordinator	6.16	4.39	5.00	7.27	4.64	6.00	7.85	4.76	7.50	21.40	14.07	19.50
H31	conce_adv_sub	0.56	0.94	0.00	0.48	0.93	0.00	0.53	0.75	0.00	3.25	2.49	3.00
H32	coconditional.adverbial.subordinator	2.70	2.24	2.00	3.43	2.35	3.00	4.88	3.46	4.50	15.75	9.00	14.00
H33	other_adverbial_subordinator	2.88	2.72	2.00	4.05	3.07	4.00	4.65	3.59	4.00	12.90	7.09	11.50
I34	prepo	76.34	19.19	76.00	84.77	22.64	80.50	100.00	24.98	103.00	338.25	59.29	331.50
I35	attributive.adjective	57.14	16.58	56.00	61.91	18.87	57.00	64.78	18.86	60.00	219.85	43.40	222.50
I36	predicative_adjective	8.06	4.26	8.00	8.25	3.74	8.00	10.35	4.68	10.00	35.45	15.21	36.00
I37	adverb	58.57	23.05	53.00	68.73	37.08	66.00	64.43	23.19	62.50	229.10	51.03	216.00
J38	conjunct	1.05	1.49	1.00	1.38	1.61	1.00	1.78	2.19	1.00	2.70	2.56	2.00
J39	downtoner	1.87	1.65	1.00	1.80	1.57	2.00	2.28	2.00	2.00	6.35	4.31	6.00
J40	hedge	4.97	4.55	4.00	5.75	4.69	5.00	5.95	4.30	5.00	25.00	16.53	20.50
J41	amp	8.61	7.17	8.00	9.63	7.78	8.00	6.80	6.32	5.50	16.70	9.67	16.50
J42	emphatic	15.48	8.40	15.00	20.41	10.81	18.00	24.30	12.88	23.00	100.60	32.39	100.00
J43	discourse_particle	1.12	1.64	1.00	1.07	1.16	1.00	0.88	1.22	0.00	2.65	2.91	1.00
K44	possibility.modal	9.61	4.90	9.00	10.38	5.89	9.00	11.60	4.42	10.50	28.25	12.88	27.00
K45	necessity.modal	0.92	1.05	1.00	1.77	2.08	1.00	1.70	2.02	1.00	2.60	1.96	2.50
K46	predictive.modal	4.25	2.80	4.00	5.25	3.65	4.00	7.18	5.46	6.00	30.75	11.60	31.50
L47	public.verb	4.60	3.35	4.00	5.04	3.17	5.00	5.93	5.16	5.00	12.10	8.41	10.00
L48	private.verb	25.86	13.09	25.00	29.29	19.55	24.00	31.18	11.05	30.00	92.35	26.86	87.00

(Table 20 continues).

(Table 20 continued).

		Level 7			Level 8			Level 9			NS		
		<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
L49	suasive.verb	2.71	2.34	2.00	2.43	1.82	2.00	2.20	1.84	2.00	5.35	3.17	4.50
L50	seem.appear	0.82	1.38	0.00	0.68	1.01	0.00	0.75	1.19	0.00	3.25	3.09	2.00
M51	contraction	37.27	11.53	37.00	40.25	13.28	40.00	49.35	17.66	48.00	190.80	54.42	190.00
M52	stranded.preposition	2.16	1.95	2.00	2.18	1.74	2.00	2.95	1.95	3.00	15.30	5.18	14.00
M53	split_infinitive	0.16	0.40	0.00	0.11	0.37	0.00	0.35	0.62	0.00	0.95	1.39	1.00
M54	split_auxiliary	1.87	1.67	2.00	2.46	1.81	2.00	2.95	2.75	2.00	10.00	4.41	9.50
N55	phrasal_coordination	4.21	2.37	4.00	3.80	2.50	3.00	4.90	3.02	4.00	20.00	7.65	22.00
N56	ind_clause_coordination	27.51	10.81	27.00	31.46	8.73	32.00	33.33	8.78	32.50	103.80	45.20	95.00
O57	synt_negation	0.73	0.95	0.00	0.77	1.19	0.00	0.75	1.06	0.00	3.60	2.56	3.00
O58	analytic_negation	19.36	6.98	18.00	19.96	6.36	20.00	25.15	7.24	24.50	57.45	20.24	55.00

Note. NS = Native speaker of English.

### Interpreting the Box-and-Whisker Plots

The box-and-whisker plots show how each linguistic feature was distributed across the oral proficiency groups. Box-and-whisker plots for all the features are provided in Appendix E. Among these box-and-whisker plots, *second person pronoun*, which includes items such as *you, your, yours, yourself, yourselves*, is used as an example to explain how to interpret the plots (Figure 3). The confidence intervals across the levels are provided in Table 21.

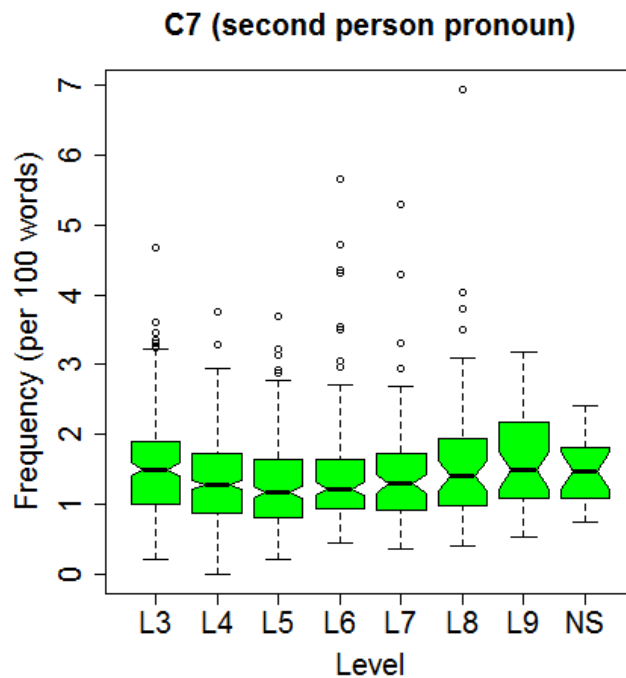


Figure 3. Box-and-whisker plots for C7 (second person pronoun).

Table 21. *Confidence Intervals Across the Levels for C7 (Second Person Pronoun)*

		L3	L4	L5	L6	L7	L8	L9	NS
C7	Upper	1.59	1.33	1.26	1.32	1.44	1.61	1.75	1.74
	Lower	1.40	1.21	1.09	1.12	1.15	1.20	1.22	1.22

Note. NS = Native speaker of English.

The vertical axis on the left-hand side of the box-and-whisker plots shows the frequency of occurrence of the grammatical feature, normalized to frequency per 100 words. This is a reasonable proportion to normalize to because most learners produced at least four hundred words in their tasks (mean tokens of the lowest oral proficiency level learners, L3 is 429.51). The horizontal axis shows the oral proficiency level of each of the eight plots, from L3 to NS. In each figure, each level has a separate box plot (eight per figure). The bottom and top line of each box shows the 25th and the 75th percentiles in the distribution of frequencies across the learners within that proficiency level. The thick line in the middle of the box indicates the 50th percentile - the median. The vertical whiskers above and below the boxes show the spread of the data. The upper whisker is at a point 1.5 times the upper quartile range above the box and the lower whisker is at a point 1.5 times the lower quartile range below the box. These vertical whiskers have a maximum length that is equal to 1.5 times the height of the box. Outliers here are defined as values that are outside the length of these vertical whiskers.

Each small circle represents one person. All of the groups except L9 and NS have outliers showing more than usual use of *second person pronoun*. Here is an example from the transcripts of an L6 learner using *second person pronoun*, “So, if *you* try to find any particular kind of thing, or if *you* know the shop, well, it could be very useful (6\_0328.txt).” A few learners (see L8 in the plot especially) had very high numbers of

*second person pronoun*, with the learner, who is the highest outlier in L8 (8\_1186.txt) using it around 7 times per 100 words. This L8 learner used the phrase “you know” very frequently during the performance as in the following examples, “*You know*, because, *you know*, I spent three years already so I just want to move another city right now.” “Because *you know*, we are paying for the apartment, *you know*.” “*You know?*” In L6 there are 8 outliers out of a total of 130 test-takers. Each small circle represents one person and each oral proficiency level involved between the lowest 40 (L9) to the highest 482 (L4) people. As shown in these examples, the box-and-whisker plots can clearly show the numbers of outliers in a figure, so it is useful in seeing the language use variation of individual test-takers.

The notches in the boxplots display the confidence interval around the median, and when the notches of boxes overlap there is ‘strong evidence’ (95% confidence) that their medians significantly match (Chambers et al, 1983). Table 21, for example, shows whether the confidence intervals overlap or not. The notches between L3 and L4 in Figure 3 do not clearly show whether there is an overlap or not. However, the confidence intervals in Table 21 (L3 Lower: 1.40 and L4 upper: 1.33) clearly show that there is no overlap between them. Except for L3 and L4, the overlaps of the notches across other oral proficiency levels indicate that there are significant similarities between the groups. Thus, *second person pronoun* is a linguistic feature that shows no frequency change pattern except for L3 and L4.

Additionally, box-and-whisker plots can provide information regarding dispersion. The example of *downtoner*, which include items such as *barely*, *nearly*, *slightly*, is used

to explain this point (Figure 4). The percentages of people using *downtoner* across the levels are L3: 55%, L4: 81%, L5: 92%, L6: 96%, L7: 97%, L8: 96%, L9: 93%, and NS: 100%. If 55 out of every 100 test-takers used *downtoner* at least once, it is calculated as 55%. The median line of the L3 box around the zero line shows that over 50% test-takers did not use the item *downtoner*.

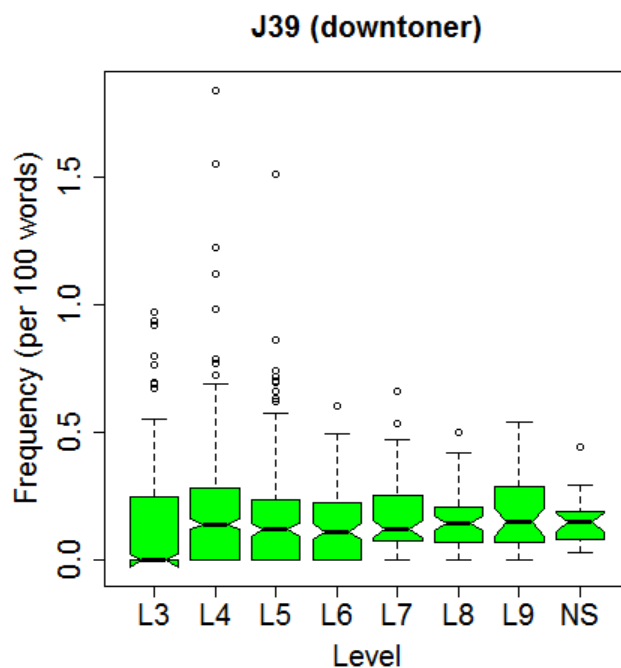


Figure 4. Box-and-whisker plots for J39 (downtoner).

It might be that lower-proficiency learners have to concentrate on the message of their speech rather than downtone their speech. The number of people using *downtoner* tended to increase beyond the lowest oral proficiency levels showing that while these are not distinctive features of oral proficiency development, nonetheless their use increases as oral proficiency develops. In fact, as in the example of *downtoner*, the

box-and-whisker plots illustrate the distribution of individual test-takers, which is summarized into five points, 0%, 25%, 50%, 75%, and 100%. The upper whisker is at a point 1.5 times the upper quartile range above the box and the lower whisker is at a point 1.5 times the lower quartile range below the box.

When interpreting the box plots, it is also necessary to consider the scale of frequencies on the vertical axis as well as the relative shapes of the eight plots.

Box-and-whisker plots for *downtoner* can be also used an example of how a low frequency feature is shown in the box-and-whisker plots. Where the frequency scale is using very small numbers, say less than one occurrence per 100 words, as in this example, we should not read too much into the differences between the plots because the differences are based on very few occurrences of the feature in each learner's oral production.

### **Results of Box-and-Whisker Plots Analysis**

In this section, I answer the first research question: what linguistic features characterize different English oral proficiency groups of Japanese learners? In order to answer this research question I present box-and-whisker plots which show the frequency change patterns across the oral proficiency levels. Using box-and-whisker plots, linguistic features can be categorized into one of the following groups: (a) falling frequency change patterns, (b) rising frequency change patterns, and (c) combination of rising or falling and plateauing change patterns. These patterns are interpreted visually in terms of the variance of individual language use (i.e., outliers, length of whiskers and size of the

boxes). Above all, the confidence intervals are used as an objective way of seeing whether differences between groups are statistically significant or not. Each frequency change pattern is presented in the following sections.

### Falling Frequency Change Pattern

There are very few plots where low proficiency learners score higher than higher proficiency learners and native speakers. The use of *noun* is a striking example where the medians clearly drop from L3 to L6 (Figure 5). The accompanied Table 22 shows confidence intervals across the levels.

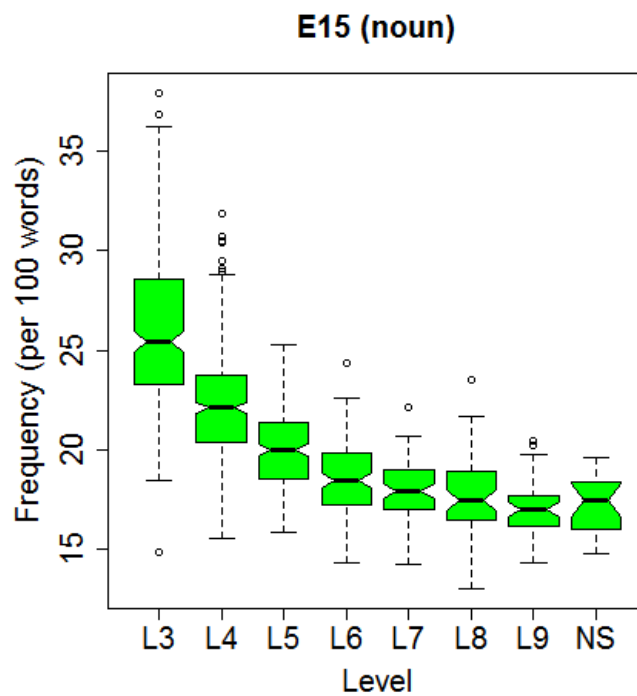


Figure 5. Box-and-whisker plots for E15 (noun).

Table 22. *Confidence Intervals Across the Levels for E15 (Noun)*

		L3	L4	L5	L6	L7	L8	L9	NS
E15	Upper	25.98	22.35	20.29	18.80	18.30	17.98	17.35	18.32
	Lower	24.86	21.87	19.71	18.10	17.56	16.95	16.60	16.65

Note. NS = Native speaker of English.

Looking across the box-and-whisker plots, there is no overlap among the notches of the boxes. The confidence intervals, which indicate statistically significant differences, support this drop (L3 lower: 24.86 and L4 upper: 22.35, L4 lower: 21.87 and L5 upper: 20.29, L5 lower: 19.71 and L6 upper: 18.80). This falling frequency change pattern begins with a high frequency with medians of 25.42 (L3) occurrences per 100 words. The box-and-whisker plot of L3 indicates that over 50% of L3 test-takers use *noun* 25 times in 100-words performance. Additionally, the long whiskers and large boxes of L3 and L4 show how variously *noun* is used by lower-level learners. The following examples show that *noun* is a distinctive feature in the performance of L3 learners (e.g., “*rice ball.*” “*red ball* and” “*in Japanese umeboshi*” “*Yeah. Convenient store.*” “*Yes. with my wife.*” 3\_0459.txt). These examples show how L3 learners list up *noun* to continue their conversation. However, in contrast to this extremely high dependency on *noun*, the decrease in the use of *noun* is possibly caused by a greater use of pronouns, coordinated predicates, and non-finite subjectless clauses. *Noun* is the most frequently used linguistic feature in this study, and it is used by 100% of test-takers in all oral proficiency groups. This falling frequency change pattern is a sign of increasing oral proficiency, so that *noun* can be considered as a linguistic feature that characterizes different English oral proficiency groups of Japanese learners.

## Rising Frequency Change Pattern

There are two box-and-whisker plots that clearly show rising frequency change patterns. I begin by presenting *emphatic* (Figure 6). It includes items such as *a lot, for sure, really* when they are used to emphasize ideas. Emphatics occur with a moderate frequency with medians ranging from 0.27 (L3) to 2.41 (NS) occurrences per 100 words.

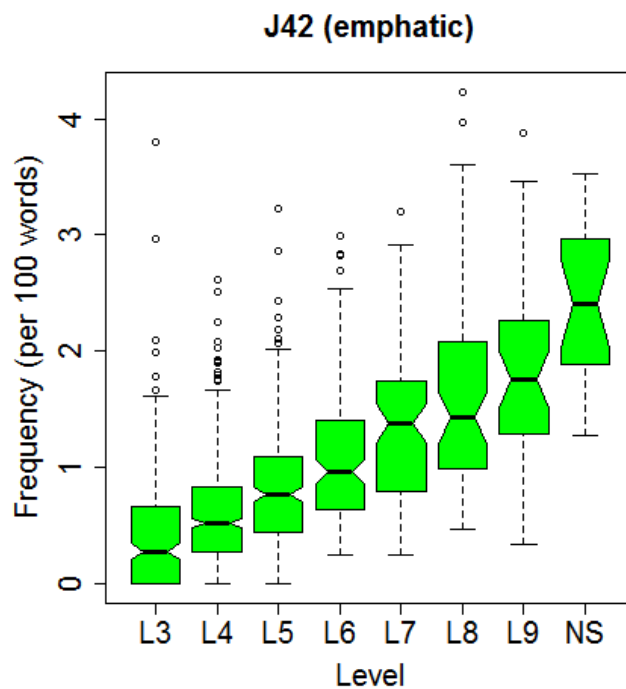


Figure 6. Box-and-whisker plots for J42 (emphatic).

Low proficiency learners (L3) use them much less often than learners at higher proficiency levels. The L3 box sits on the bottom line of the figure indicating that over 25% of L3 test-takers did not use emphatics. However, note that there is a regular increase in use through the oral proficiency levels. Looking across the plots there is no overlap across the notches of the boxes from L3 to L7 learners and between L9 and NS.

The statistically significant differences are shown by confidence intervals in Table 23 (L3 upper: 0.34 and L4 lower: 0.47, L4 upper: 0.55 and L5 lower: 0.70, L5 upper: 0.83 and L6 lower 0.85, L6 upper: 1.06 and L7 lower: 1.20, L9 upper: 2.00 and NS lower: 2.02).

Table 23. *Confidence Intervals Across the Levels for J42 (Emphatic)*

		L3	L4	L5	L6	L7	L8	L9	NS
J42	Upper	0.34	0.55	0.83	1.06	1.54	1.66	2.00	2.79
	Lower	0.20	0.47	0.70	0.85	1.20	1.19	1.51	2.02

Note. NS = Native speaker of English.

The increasing use of *emphatic* is a sign of increasing oral proficiency, and it can be also considered as a distinctive feature for native speakers (e.g., “But it’s been a good experience *for sure*.” NS\_0005.txt). Note also that some lower oral proficiency learners (the outliers at the top of the plot) used them with a roughly similar frequency to native speakers. One of the L3 learners (3\_1096.txt) used the word “so” frequently for emphasis as in the following examples, “But *so* many people.” “It’s *so* delicious.” “It’s *so* warm.” “It’s *so* cheap.” “It’s *so* difficult to explain.”

*Pronoun it* also displays the rising frequency change pattern (Figure 7). The box-and-whisker plots demonstrate it well. There are no overlaps across the notches of the boxes from L3 to L5 learners, L6 to L7 learners, and L9 to native speakers of English. The confidence intervals in Table 24 indicate statistically significant differences between these proficiency groups (L3 upper: 0.85 and L4 lower: 1.01, L4 upper: 1.14 and L5 lower: 1.30, L6 upper: 1.60 and L7 lower: 1.85, L9 upper: 2.26 and NS lower: 2.44).

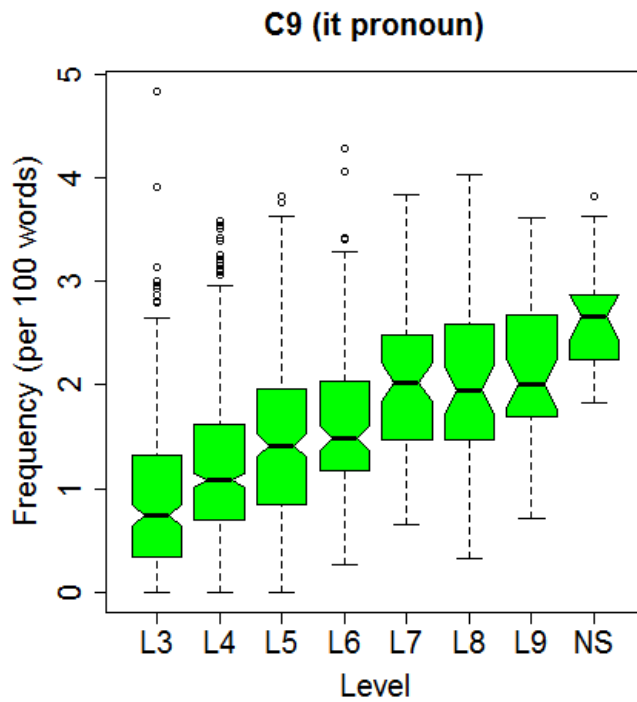


Figure 7. Box-and-whisker plots for C9 (pronoun *it*).

Table 24. Confidence Intervals Across the Levels for C9 (Pronoun *it*)

		L3	L4	L5	L6	L7	L8	L9	NS
C9	Upper	0.85	1.14	1.53	1.60	2.21	2.19	2.26	2.88
	Lower	0.64	1.01	1.30	1.36	1.85	1.72	1.76	2.44

Note. NS = Native speaker of English.

Observing example sentences, native speakers of English can continuously talk about a particular topic (e.g., I don't like cats at all. And to begin with, *it*, like, threw up all over, which doesn't even make me like *it* more. I hate cats. Anyway, but she really loves the cat, so I have to, like, pretend I like *it*. And she always like makes faces at *it*, and I'm like, "Oh *it*'s really cute." But I really don't like *it* at all. And, yeah, the other day, *it* just threw up all over and I was like "Oh good lord." Now, I try to keep my door closed

so *it* doesn't go in and like throw up on all my things. But, yeah, yeah, I don't really like the cat. And *it* jumps on the CD player and skips my CDs whenever I'm playing CDs." NS\_00006). Yet, non-native speakers cannot continue to talk without an interruption, so the frequency of *pronoun it* can be lower than that of native speakers. However, even among non-native speakers, there is a huge gap between L6 (intermediate mid) and L7 (intermediate mid-plus). The following examples of L6 learners show how awkwardly *pronoun it* is used in speech (e.g., "I bought a new pen in your shop, but, unfortunately, I found some trouble on *it*. So I'd like to propose you to change this one to another new one that works well. Writing is not so good. Ink is sometimes disappear, the line is dotted like this here. When I tried *it* in your shop, that's OK. But I found *it* when I came back home and tried to write some letters" 6\_0227.txt). L7 learners however can continue to talk about the same topic by using the pronoun *it* properly as in the following examples (e.g., "Yes, *it's* like that. *it's* in the country side and there is lots of fields surrounding my house and the houses are each house are quite big and they all have a huge yard and they usually have pets, like dog or cats or that kind of. And lot of children live there. And excuse me. We have lot of parks and there is a big river called XXX06 River near our house and the scenery is very good. So *it's* a beautiful place" 7\_0037.txt).

The increasing use of *pronoun it* can be another sign of increasing oral proficiency. Additionally, *pronoun it* can be used to distinguish language use between native and non-native speakers of English. The use of the pronoun *it* involves the use of reference to connect to previous nouns, phrases or clauses. It might be revealing in future studies to distinguish what *it* refers to, but this is still beyond the limits of

computer-based analysis. To sum up, *emphatic* and *pronoun it* show an increasing frequency change pattern of use across the oral proficiency levels. Thus, these two features can characterize different English oral proficiency groups of Japanese learners, and they are both distinctive features for native speakers of English.

### **Other Frequency Change Patterns**

There are four frequency change groups other than falling and rising patterns: (a) rising and plateau, (b) plateau and falling, (c) falling and plateau, and (d) plateau and rising change patterns. Some of features in this study can be categorized into one of the following groups as follows: (a) rise-plateau (*infinitive, possibility modal, agentless passive, adverb, causative adverbial subordinator, hedge, contraction, indefinite pronoun, demonstrative pronoun, private verb*, and possibly *past tense*), (b) plateau-fall (*amplifier*), (c) fall-plateau (*phrasal coordination, present tense*), and (d) plateau-rise (possibly *perfect aspect*).

As shown in Figure 8, *infinitive* and *possibility modal* are not part of simple and clearly shown rising or falling frequency change patterns, but they have a kind of rising plateau frequency change patterns. Note that these two items are less frequently used by lower-level learners and native speakers of English when compared with the others. Checking the confidence intervals in Table 25, there is no overlap in the notches of boxes across the following levels: *infinitive* (L3 upper: 1.10 and L4 lower: 1.54, L9 lower: 1.69 and NS upper: 1.63) and *possibility modal* (L3 upper: 0.30 and L4 lower: 0.40, L4 upper: 0.49 and L5 lower: 0.64, L9 lower: 0.71 and NS upper: 0.70).

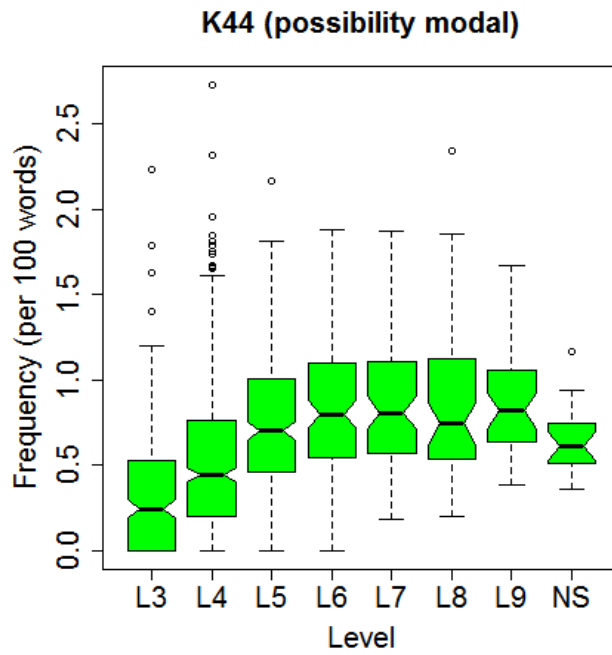
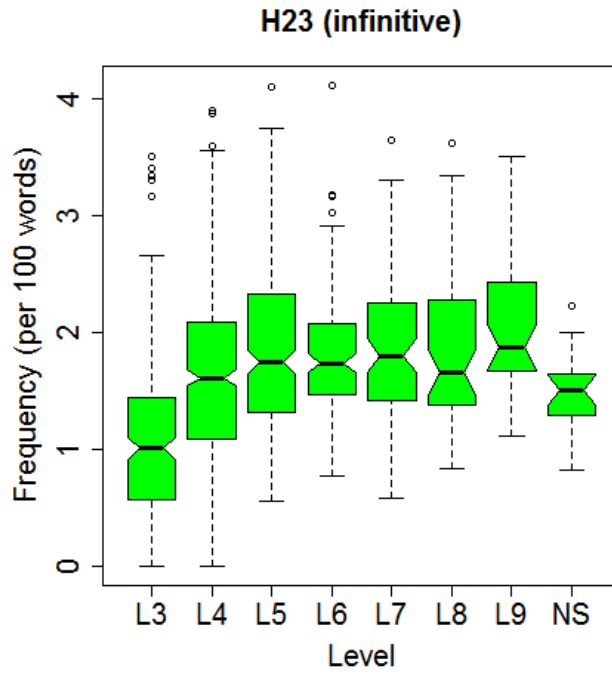


Figure 8. Box-and-whisker plots for H23 (infinitive) and K44 (possibility modal).

Table 25. *Confidence Intervals Across the Levels for H23 (Infinitive) and K44 (Possibility Modal)*

		L3	L4	L5	L6	L7	L8	L9	NS
H23	Upper	1.10	1.68	1.85	1.82	1.95	1.85	2.06	1.63
	Lower	0.91	1.54	1.64	1.65	1.65	1.47	1.69	1.37
K44	Upper	0.30	0.49	0.76	0.88	0.90	0.86	0.92	0.70
	Lower	0.19	0.40	0.64	0.72	0.71	0.62	0.71	0.53

*Note.* NS = Native speaker of English.

Novice level learners (L3) can simply connect words or phrases in the form of phrasal coordination (e.g., “raw fish *and* fried fish *and* steaming fish” 3\_1096.txt), but cannot construct complete sentences. Thus, they have the possibility of using *infinitive* less frequently than the other level learners (e.g., “I want *to* get to Tokyo Station as early as possible.” 3\_1160.txt). It is unlikely to be for the same reason, but native speakers of English use *infinitive* less frequently, too (e.g., “So I try *to* come up with things *to* do, but when there’s really nothing *to* do, I play Gameboy in my room by myself because I have no TV or computer.” NS\_0004.txt). The compact size of the box-and-whisker plot of native speakers compared to the others indicates that they tend to use *infinitive* in a more homogeneous way.

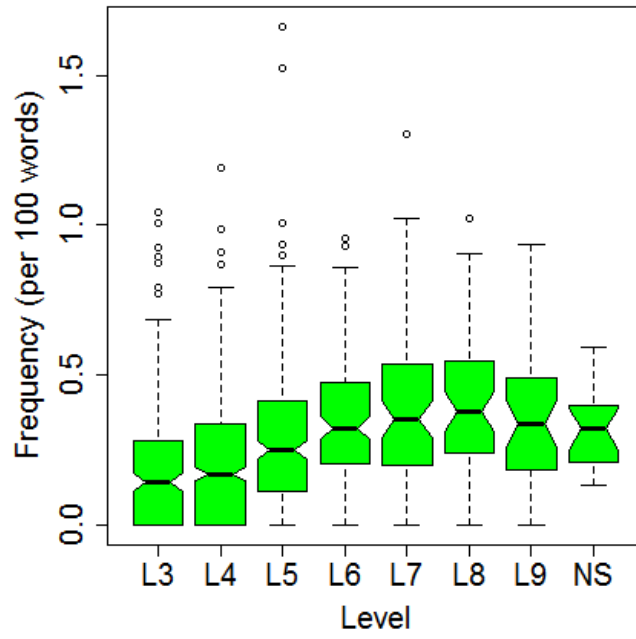
L3 and L4 learners tend to use *possibility modal* less frequently than other learners (“We *can*’t understand his statement.” 3\_0427.txt). In a similar way, native speakers of English use *possibility modal* less frequently than L5, L6, L7, L8 and L9 learners of English. This difference between native and non-native speakers of English learners might be caused by the language use tendency of Japanese EFL learners, who tend to use *possibility modal* frequently as in the following examples (e.g., “I *can*’t explain. *Pachinko*, they have some curious. How *can* I say? I *can*’t explain what is

attractive *pachinko* for me.” 4\_0801.txt). In addition, the compact size of the box-and-whisker plot of native speakers indicates a more homogeneous use of the feature. These rising and falling plateauing frequency change patterns cannot characterize all the English oral proficiency groups, but at least they can be a clear sign for L3 learners and native speakers of English. What is more, they show that language features appearing in the oral performance of highest proficiency learners do not match with those of native speakers of English.

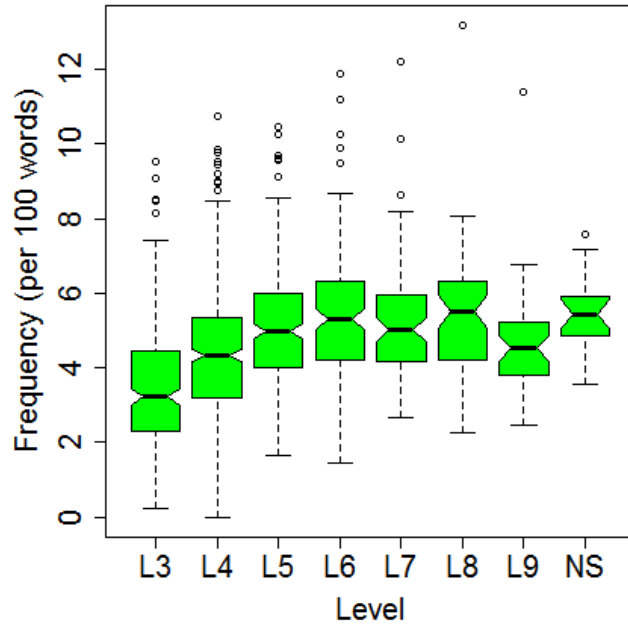
In addition to *infinitive* and *possibility modal*, the other nine linguistic features, *agentless passive*, *adverb*, *causative adverbial subordinator*, *hedge*, *contraction*, *indefinite pronoun*, *demonstrative pronoun*, *private verb*, and possibly *past tense*, also have a rise-plateau frequency change pattern (see Figure 9).

As can be seen in these box-and-whisker plots, frequencies increase as the oral proficiency level rises and reach a turning point between the rise and plateau. These relationships between oral proficiency levels and some of the variables are weaker than the previous relationships (i.e., *noun*, *emphatic*, *pronoun it*), but they can be also considered as features that characterize the change of oral proficiency levels. For example, *indefinite pronoun* is frequently used by higher-level learners and native speakers of English. This language use tendency suggests that lower-level language learners tend to concentrate on personal talk using first person pronouns (e.g., “*I* like American movies.” 4\_0065.txt), but higher-level learners come to generalize their talk by using indefinite pronouns (e.g., “But *everybody* brings *something*.” 9\_1277.txt).

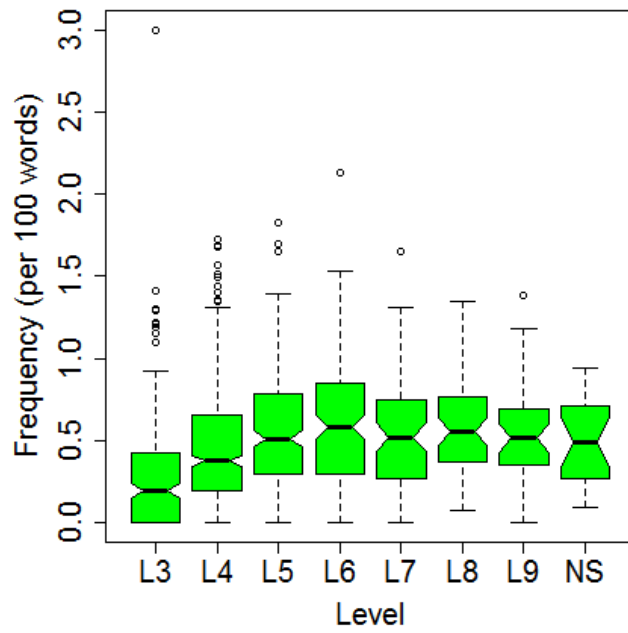
**F16 (agentless passive)**



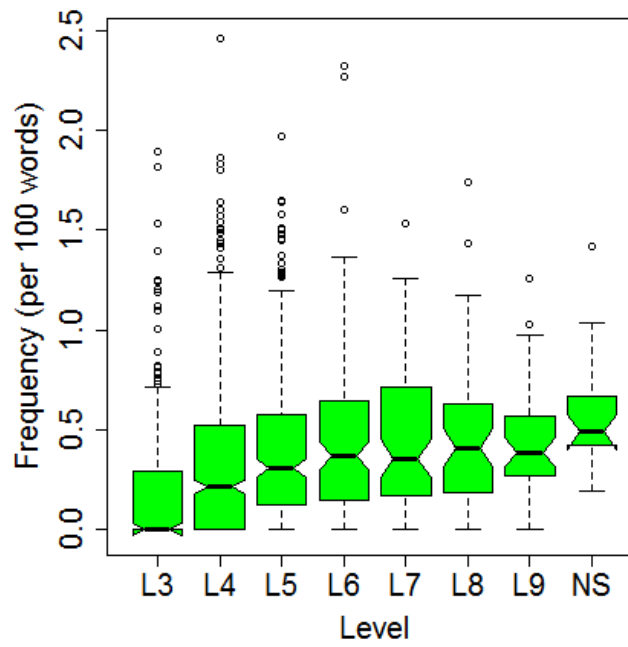
**I37 (adverb)**



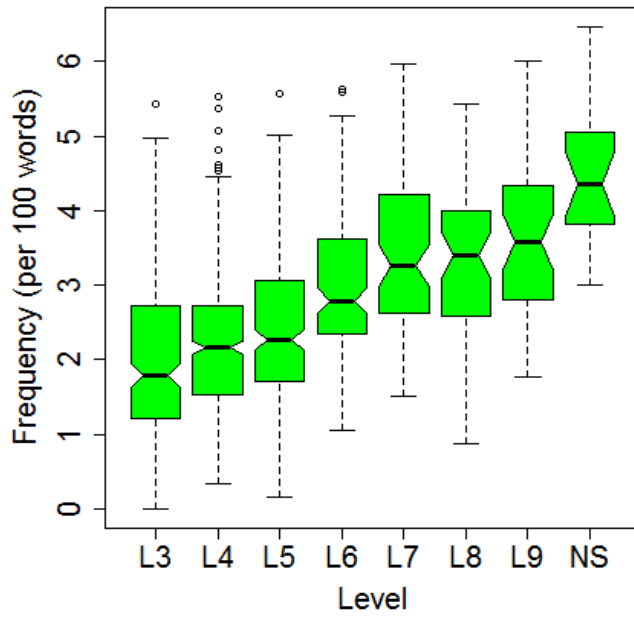
### H30 (causative adverbial subordinator)



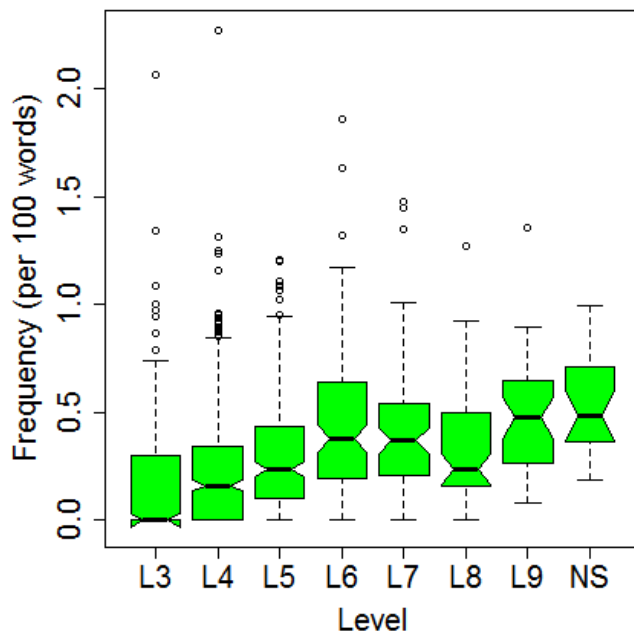
### J40 (hedge)



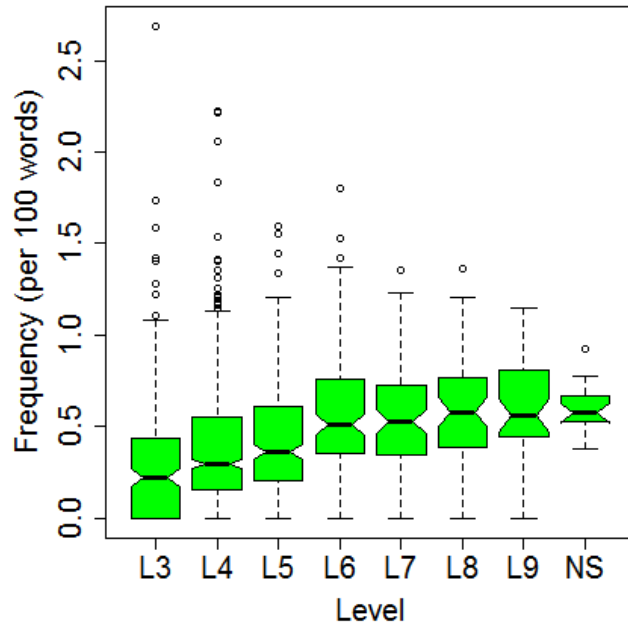
**M51 (contraction)**



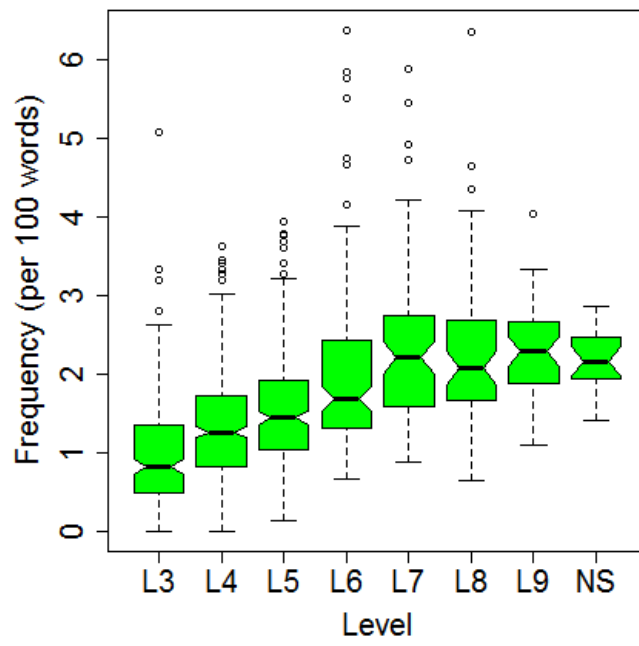
**C11 (indefinite pronoun)**



**C10 (demonstrative pronoun)**



**L48 (private verb)**



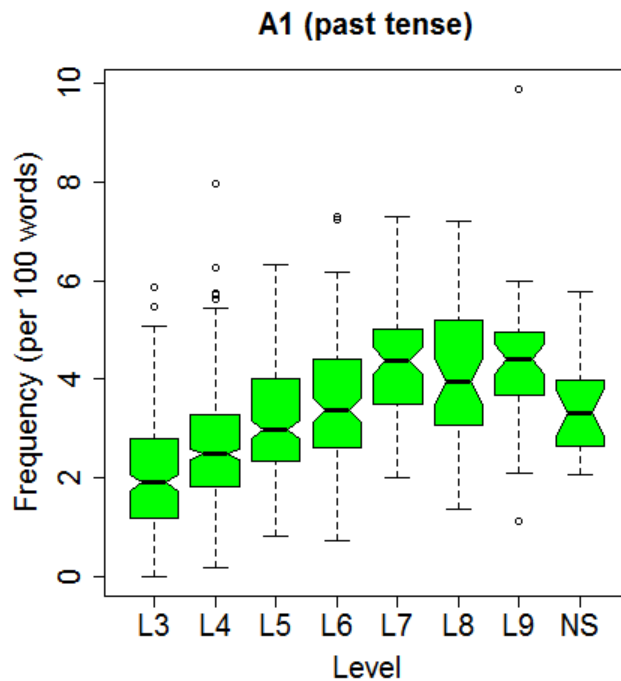


Figure 9. Box-and-whisker plots for F16 (agentless passive), I37 (adverb), H30 (causative adverbial subordinator), J40 (hedge), M51 (contraction), C11 (indefinite pronoun), C10 (demonstrative pronoun), L48 (private verb), and A1 (past tense).

Another type of combination pattern was found in this study—the plateau and falling frequency change pattern (Figure 10). *Amplifier*, which includes items such as *absolutely*, *completely*, *totally*, *very* has this pattern. However, in the *amplifier* category, *very* dominates most of the tokens (*very*: 10,131 tokens, *amplifier* except *very*: 196 tokens). Japanese EFL learners appear to have a tendency to use the word *very* as a part of chunks in “thank you *very* much” or “*very* good.” Thus, the falling frequency change pattern of *amplifier* can be considered as a decrease in the use of the word *very*. There is no significant difference between the most advanced L9 learners and native speakers of

English, but interestingly there is a huge gap between L6 (intermediate mid) and L7 (intermediate mid-plus) learners as in the case of *pronoun it*.

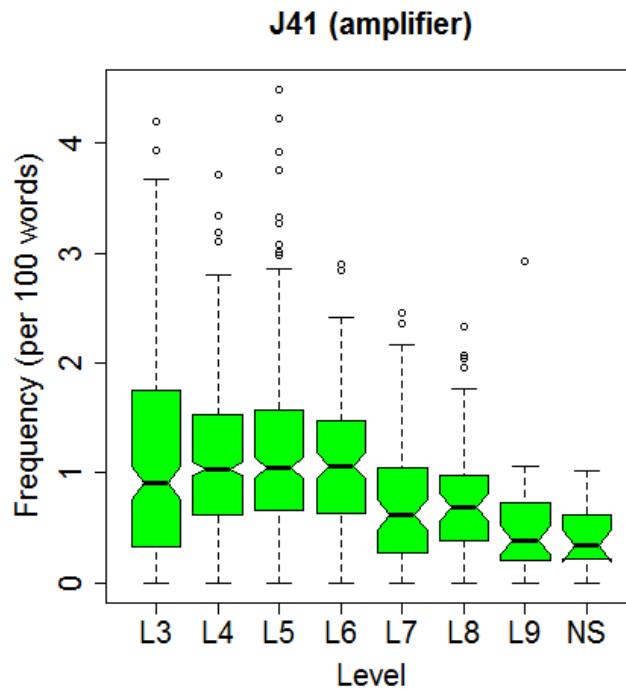


Figure 10. Box-and-whisker plots for J41 (amplifier).

There are two more frequency change groups other than simple falling and rising patterns: falling and plateau change pattern (*phrasal coordination, present tense*) and plateau and rising change patterns (*perfect aspect*). I begin by looking at the changes in *phrasal coordination* (Figure 11). *Phrasal coordination* is most frequently used by L3 learners (e.g., “raw fish *and* fried fish *and* steaming fish” “And I cooked fried egg *and* something” “speak clearly *and* loudly.” 3\_1096.txt.). There is a fall from L3 to L5/L6 and then a plateau. Additionally, there was no significant difference between the use of L9

and NS, but native speaker use of this feature was more frequent than the more advanced learner groups (e.g., “But my other classes, Japanese politics *and* Japanese economics are both in English.” “And so it was fun just to talk back *and* forth *and* mess around.” “...but they have like bowls *and* dishes *and* chopsticks *and* all these different kind of stuff.” “And there seems to be a lot of foliage, meaning bushes *and* trees *and* stuff.” “Me *and* my friends take road trips.” “Squeak *and* Flash.” “...so just makes me feel kind of calm *and* relax when I go there.” “XXX14 now is a world-wide organization dedicated to eliminating poverty housing *and* homelessness from the face of the earth by basically building houses for people.” NS\_0020.txt.). As in these examples, *phrasal coordination* is used effectively by connecting various part-of-speech words. In contrast to this falling frequency change pattern of *phrasal coordination*, the use of *conjunct* (e.g., *however*) and *independent clause coordination* (e.g., “Usually, I get up late *and* clean my room.” 5\_0003.txt) did not did not show a clear frequency change or a linear development. However, considering the phenomenon that the frequency of *phrasal coordination* decreased, learner language shows a shift from connecting phrases to sentences.

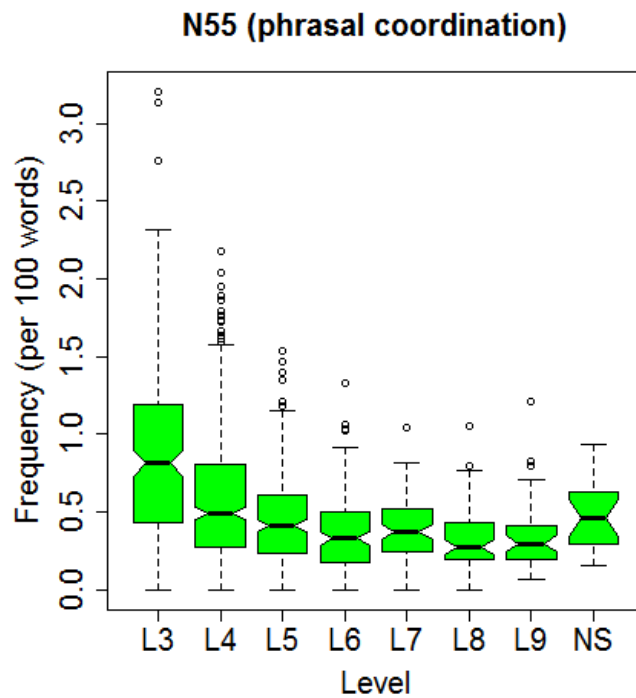


Figure 11. Box-and-whisker plots for N55 (phrasal coordination).

As can be seen from the box-and-whisker plots in Figure 12, the frequencies of *present tense* decrease as the oral proficiency level rises and reaches a turning point between the fall and plateau. On the other hand, the frequency of *perfect aspect* increases from L4 to L5 but then reaches a turning point between rise and plateau; the frequency clearly increases between L9 learners and native speakers of English.

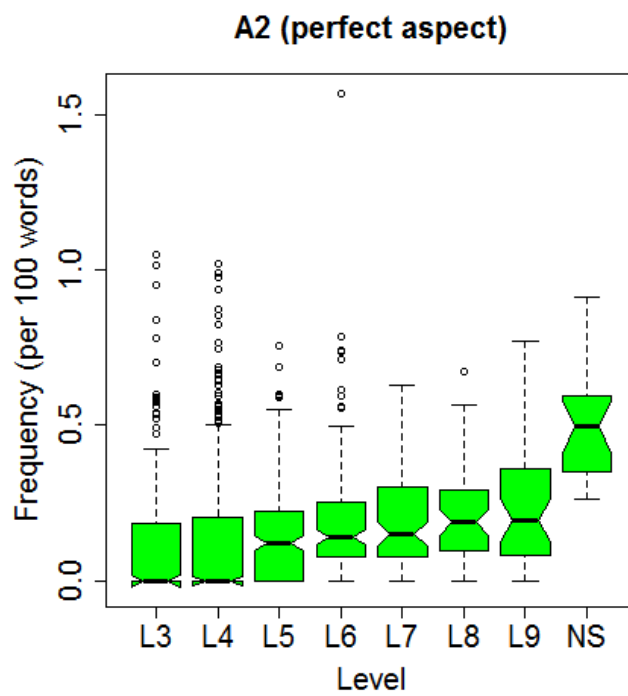
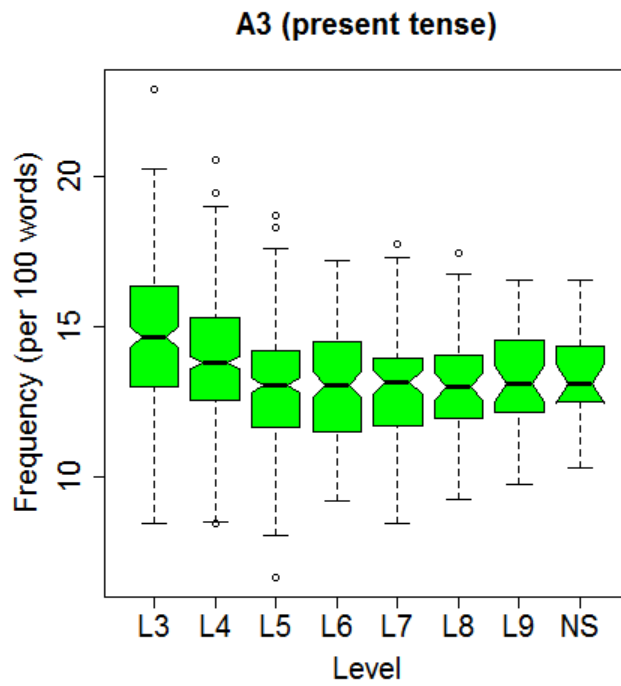


Figure 12. Box-and-whisker plots for A3 (present tense) and A2 (perfect aspect).

In general, *present tense* is necessary for constructing simple basic English sentences and for doing the tasks in the speaking test. As it is a basic essential of English, it is used with roughly equal frequency by each group and it does not distinguish any oral proficiency group (L5, L6, L7, L8, L9, NS). However, they are more frequently used by lower-level learners (L3 and L4), who might not have a good command of other tenses, such as past and future tense. Thus, in contrast to *present tense*, the box-and-whisker plot of *past tense* presents a rising pattern with a lowish median for L8 which however is still within the ranges of the other learners' groups (Figure 9). In general, the present grammatical analyses confirm the overall ratings given by the Standard Speaking Test (SST) test raters. Recall, as mentioned in Chapter 3, Methods, none of the grammar features, except for *present tense* and *past tense* are included in the criteria of SST. The frequency of *past tense*, for example, increases as the oral proficiency level rises, and this can be related to the learners' increasing ability to inflect verbs for tense. However, it is interesting to note that the use of the *past tense* is lower among native speakers of English than high proficiency learners. This could occur because native speakers of English possibly use *perfect aspect* in place of *past tense*. The box-and-whisker plot clearly shows that native speakers of English use *perfect aspect* much more often than learners at higher oral proficiency levels (Figure 12). Note that high oral proficiency learners have a tendency to use *past tense* when they talk about their experience in the past (e.g., "As I *told* you, I *went* to the United States. And, before that, I *was* attracted the American culture. And, especially, even before that, I *went* to the United States." 9\_0325.txt). On

the other hand, native speakers of English are likely to use present perfect instead of simple past tense (e.g., “I’ve *traveled* a lot on the East Coast. Mostly all the states on the East Coast. And I’ve *been* to California and my dad has friends in California and he has businesses there sometimes.” NS\_0006.txt). Consequently, *perfect aspect* can be used as a sign to distinguish native and non-native speakers of English.

The patterns shown in the box-plot analysis provide useful insights into signs of increasing proficiency. They also show that an increase in the range of grammatical features used tends to be a sign of greater proficiency. Now, having looked at details of grammatical feature use, let us look at the broader picture.

### **Results of Correspondence Analysis**

Correspondence analysis is an exploratory data analysis technique that is useful in analyzing a two-way contingency table. Unlike hypothesis testing statistical techniques (e.g., chi-square, log-likelihood) that verify an a priori hypothesis regarding relations between variables, it is designed to identify relations between a large number of variables without setting an a priori hypothesis. It can work with a large number of data points that cannot be easily handled through a visual inspection or a series of pairwise comparisons of variables.

This statistical technique requires a data matrix with non-negative entries and with the data rated on the same scale. Table 26 and Table 27 are the two-way contingency tables that contain the data used in the present study. The label for each linguistic feature is based on the grammatical classification used in Biber (1988). Table

26 shows normalized frequency counts in one million words of 58 grammatical features for eight proficiency levels. Table 27 shows raw frequency counts of 58 grammatical features for eight proficiency levels. The columns are oral proficiency levels and the rows are grammatical features. The data at the intersections of each row and column are frequency data (How often a particular grammatical feature occurs at a certain proficiency level).

The raw data and normalized data look different because the raw data preserve information about how many words the speakers in each group actually produce but the standardized data remove it. However, correspondence analysis begins by normalizing all the data, so, in order to avoid normalizing the data twice, raw frequency counts were used.

In order to better represent the performance of the whole group, raw frequency counts of medians were multiplied by the number of test-takers at each oral proficiency level. Medians represent only one subject in each group, but they do not provide much confidence about the group coordinates. Thus, the group medians were multiplied by the number of test-takers in each group, and this would make it equivalent to the group sum totals. The reason for focusing on medians is that there are obviously non-normal distributions in some groups, and so medians can better measure the central tendency.

It is not worth keeping all the 58 linguistic features in the correspondence analysis, because they do not have good enough quality data to be included in the analysis. Some of the linguistic features, *by passive* (F17), *that adjective compound*

Table 26. Normalized Frequency Counts in One Million Words of 58 Grammatical Features for Eight Proficiency Levels

Linguistic features	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	NS
A1 past tense	21,080	25,847	31,679	35,010	42,380	40,050	43,553	34,940
A2 perfect aspect	1,038	1,329	1,509	1,860	1,897	1,999	2,353	4,919
A3 present tense	144,066	138,752	128,852	130,848	130,382	130,758	132,220	132,729
B4 place adverb	3,681	3,497	3,700	3,352	3,525	3,049	3,953	5,237
B5 time adverbial	6,628	6,929	7,312	6,925	6,581	5,734	5,368	4,223
C6 first person pronoun	88,116	85,203	84,759	84,153	87,944	82,216	82,344	70,824
C7 second person pronoun	14,913	13,123	12,595	14,830	15,024	17,713	16,380	15,087
C8 third person pronoun	22,695	23,799	24,558	24,312	22,519	23,053	26,474	23,592
C9 pronoun <i>it</i>	9,428	12,403	14,815	16,208	19,779	19,595	21,399	26,966
C10 demonstrative pronoun	2,957	3,918	4,465	5,823	5,644	6,055	6,140	5,969
C11 indefinite pronoun	1,919	2,356	2,985	4,530	4,462	3,283	4,890	5,391
C12 proverb <i>do</i>	2,255	2,398	2,475	2,908	4,356	3,983	5,166	5,473
D13 direct WH-question	2,957	2,515	1,946	1,867	1,183	1,517	1,471	1,970
E14 nominalization	6,901	8,952	9,405	9,535	8,256	8,112	7,354	8,245
E15 total other nouns (except for nominalization)	258,537	220,406	199,767	184,775	179,390	176,761	170,350	172,010
F16 agentless passive	1,898	2,065	2,901	3,444	3,829	4,041	3,640	3,197
F17 <i>by</i> passive	84	139	152	321	316	248	202	224
G18 <i>be</i> main verb	13,927	13,379	12,654	12,772	12,553	12,095	13,145	13,129
G19 existential <i>there</i>	4,520	4,978	4,626	4,102	3,818	4,027	4,284	5,497

(Table 26 continues).

(Table 26 continued.)

Linguistic features	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	NS
H20 <i>that</i> verb complement	220	386	647	1,002	1,499	2,043	1,838	1,427
H21 <i>that</i> adjective complement	10	36	93	130	269	292	294	318
H22 WH-clause	378	567	833	926	1,464	1,284	1,526	1,533
H23 infinitive ( <i>to</i> -clause)	11,001	16,490	18,623	17,968	18,561	18,077	20,370	14,851
H24 past participial postnominal (reduced relative) clause	189	298	284	406	246	277	257	366
H25 <i>that</i> relative in subject position	94	133	108	115	316	248	735	1,333
H26 <i>that</i> relative in object position	126	285	289	321	749	948	1,452	1,911
H27 WH relative in subject position	136	525	936	1,301	1,265	1,386	1,949	1,380
H28 WH relative in object position	52	143	191	306	176	263	239	260
H29 WH relative with fronted preposition	0	16	29	54	59	73	110	35
H30 causative adverbial subordinator	2,601	4,567	5,646	5,946	5,551	5,938	5,773	5,049
H31 concessive adverbial subordinator	63	58	123	130	504	394	386	767
H32 conditional adverbial subordinator	598	1,280	1,921	2,303	2,436	2,801	3,585	3,716
H33 other adverbial subordinator	514	788	1,098	1,967	2,600	3,312	3,419	3,043
I34 preposition	58,153	66,768	69,768	67,632	68,833	69,260	73,538	79,800
I35 attributive adjective	54,451	55,081	54,223	51,723	51,525	50,584	47,634	51,867
I36 predicative adjective	7,614	7,347	7,165	7,163	7,272	6,741	7,611	8,363
I37 adverb	34,776	45,199	51,424	54,692	52,813	56,158	47,377	54,050
J38 conjunct	787	1,144	1,372	1,201	949	1,123	1,305	637

(Table 26 continues).

(Table 26 continued.)

Linguistic features	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	NS
J39 downtoner	1,405	1,789	1,637	1,423	1,686	1,474	1,673	1,498
J40 hedge	2,370	3,494	4,264	4,813	4,485	4,698	4,375	5,898
J41 amplifier	11,914	11,366	11,958	10,989	7,764	7,864	5,001	3,940
J42 emphatic	4,698	6,271	8,621	11,050	13,959	16,677	17,870	23,734
J43 discourse particle	1,196	1,734	1,774	1,821	1,007	875	643	625
K44 possibility modal	3,408	5,448	7,704	8,326	8,666	8,477	8,530	6,665
K45 necessity modal	587	651	877	1,232	831	1,444	1,250	613
K46 predictive modal	3,618	4,926	4,768	4,255	3,829	4,290	5,276	7,255
L47 public verb	2,664	3,043	3,754	4,676	4,145	4,114	4,357	2,855
L48 private verb	9,911	13,457	15,418	20,485	23,315	23,928	22,925	21,787
L49 suasive verb	1,164	2,473	2,504	2,724	2,447	1,984	1,618	1,262
L50 seem and appear	157	263	392	750	738	554	552	767
M51 contraction	19,779	22,276	24,068	29,186	33,609	32,886	36,291	45,014
M52 stranded preposition	1,898	1,381	1,421	1,661	1,944	1,780	2,169	3,610
M53 split infinitive	42	42	34	84	141	88	257	224
M54 split auxiliary	734	904	1,367	1,676	1,686	2,013	2,169	2,359
N55 phrasal coordination	8,736	5,604	4,421	3,421	3,794	3,108	3,603	4,718
N56 independent clause coordination	21,185	25,147	24,347	23,921	24,802	25,708	24,506	24,489
O57 synthetic negation	1,699	1,154	1,088	987	656	627	552	849
O58 analytic negation	9,344	10,841	11,688	15,726	17,460	16,312	18,495	13,554

Note. NS = Native speaker of English.

Table 27. Raw Frequency Counts of 58 Grammatical Features for Eight Proficiency Levels

Linguistic features	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	NS
A1 past tense	2,010	7,975	6,464	4,575	3,619	2,745	2,369	2,962
A2 perfect aspect	99	410	308	243	162	137	128	417
A3 present tense	13,737	42,811	26,292	17,099	11,134	8,962	7,192	11,252
B4 place adverb	351	1,079	755	438	301	209	215	444
B5 time adverbial	632	2,138	1,492	905	562	393	292	358
C6 first person pronoun	8,402	26,289	17,295	10,997	7,510	5,635	4,479	6,004
C7 second person pronoun	1,422	4,049	2,570	1,938	1,283	1,214	891	1,279
C8 third person pronoun	2,164	7,343	5,011	3,177	1,923	1,580	1,440	2,000
C9 pronoun <i>it</i>	899	3,827	3,023	2,118	1,689	1,343	1,164	2,286
C10 demonstrative pronoun	282	1,209	911	761	482	415	334	506
C11 indefinite pronoun	183	727	609	592	381	225	266	457
C12 proverb <i>do</i>	215	740	505	380	372	273	281	464
D13 direct WH-question	282	776	397	244	101	104	80	167
E14 nominalization	658	2,762	1,919	1,246	705	556	400	699
E15 total other nouns (except for nominalization)	24,652	68,005	40,762	24,146	15,319	12,115	9,266	14,582
F16 agentless passive	181	637	592	450	327	277	198	271
F17 <i>by</i> passive	8	43	31	42	27	17	11	19
G18 <i>be</i> main verb	1,328	4,128	2,582	1,669	1,072	829	715	1,113
G19 existential <i>there</i>	431	1,536	944	536	326	276	233	466
H20 <i>that</i> verb complement	21	119	132	131	128	140	100	121
H21 <i>that</i> adjective complement	1	36	11	19	170	17	121	23
H22 WH-clause	125	20	16	27	175	88	83	130
H23 infinitive ( <i>to</i> -clause)	1,049	5,088	3,800	2,348	1,585	1,239	1,108	1,259

(Table 27 continues).

(Table 27 continued.)

	Linguistic features	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	NS
H24	past participial postnominal (reduced relative) clause	18	92	58	53	21	19	14	31
H25	<i>that</i> relative in subject position	9	41	22	15	27	17	40	113
H26	<i>that</i> relative in object position	12	88	59	42	64	65	79	162
H27	WH relative in subject position	13	162	191	170	108	95	106	117
H28	WH relative in object position	5	44	39	40	15	18	13	22
H29	WH relative with fronted preposition	0	5	6	7	5	5	6	3
H30	causative adverbial subordinator	248	1,409	1,152	777	474	407	314	428
H31	concessive adverbial subordinator	6	18	25	17	43	27	21	65
H32	conditional adverbial subordinator	57	395	392	301	208	192	195	315
H33	other adverbial subordinator	49	243	224	257	222	227	186	258
I34	preposition	5,545	20,601	14,236	8,838	5,878	4,747	4,000	6,765
I35	attributive adjective	5,192	16,995	11,064	6,759	4,400	3,467	2,591	4,397
I36	predicative adjective	726	2,267	1,462	936	621	462	414	709
I37	adverb	3,316	13,946	10,493	7,147	4,510	3,849	2,577	4,582
J38	conjunct	75	353	280	157	81	77	71	54
J39	downtoner	134	552	334	186	144	101	91	127
J40	hedge	226	1,078	870	629	383	322	238	500
J41	amplifier	1,136	3,507	2,440	1,436	663	539	272	334
J42	emphatic	448	1,935	1,759	1,444	1,192	1,143	972	2,012
J43	discourse particle	114	535	362	238	86	60	35	53
K44	possibility modal	325	1,681	1,572	1,088	740	581	464	565
K45	necessity modal	56	201	179	161	71	99	68	52
K46	predictive modal	345	1,520	973	556	327	294	287	615

(Table 27 continues).

(Table 27 continued).

	Linguistic features	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	NS
L47	public verb	254	939	766	611	354	282	237	242
L48	private verb	945	4,152	3,146	2,677	1,991	1,640	1,247	1,847
L49	suasive verb	111	763	511	356	209	136	88	107
L50	seem and appear	15	81	80	98	63	38	30	65
M51	contraction	1,886	6,873	4,911	3,814	2,870	2,254	1,974	3,816
M52	stranded preposition	181	426	290	217	166	122	118	306
M53	split infinitive	4	13	7	11	12	6	14	19
M54	split auxiliary	70	279	279	219	144	138	118	200
N55	phrasal coordination	833	1,729	902	447	324	213	196	400
N56	independent clause coordination	2,020	7,759	4,968	3,126	2,118	1,762	1,333	2,076
O57	synthetic negation	162	356	222	129	56	43	30	72
O58	analytic negation	891	3,345	2,385	2,055	1,491	1,118	1,006	1,149

Note. NS = Native speakers of English.

(H21), *past participle postnominal clause* (H24), *that relatives in subject position* (H25), *WH relative in object position* (H28), *WH relatives with fronted preposition* (H29), *concessive adverbial subordinator* (H31), *seem appear* (L50) and *split infinitive* (M53), were deleted from the analysis. All of the medians of these linguistic features in the learner groups are zero, and they do not discriminate between the oral proficiency groups. This deletion of linguistic features can help answer the third research question more fully (Research Question 3: Is the oral production of Japanese EFL learners rich enough to display the full range of features used by Biber?). In addition to this deletion, correspondence analyses were conducted with and without the data of native speakers. The main focus of the present study is to understand the linguistic variation across different oral proficiency levels. Thus, it might not be necessary to include the data of native speakers in the analysis. This deletion will help provide a clearer answer to the first research question (Research Question 1: What linguistic features characterize different English oral proficiency groups of Japanese learners?). However, it is necessary to include the data of native speakers to answer the second research question (Research Question 2: To what degree do the language features appearing in the spoken production of high proficiency learners match with those of native speakers of English who perform the same task?), and so both analyses, with and without native speakers, were used.

As a first step in correspondence analysis, Pearson's chi-square test for independence is used to see if the null hypothesis that there is no relationship between the rows and the columns can be rejected ( $X^2 = 26427.1$ ,  $df = 288$ ,  $p\text{-value} < 2.2e-16$ ). As the null hypothesis is rejected, the result indicates that there is a relationship between

variables (linguistic features) and cases (oral proficiency levels). Accordingly, as a next step the data in the contingency table are represented by a scatterplot. Figure 13 plots the relationships between the cases in a visual display, and Table 28 shows the position of each oral proficiency level in this plot. The coordinate is ranked in the score of Dimension 1.

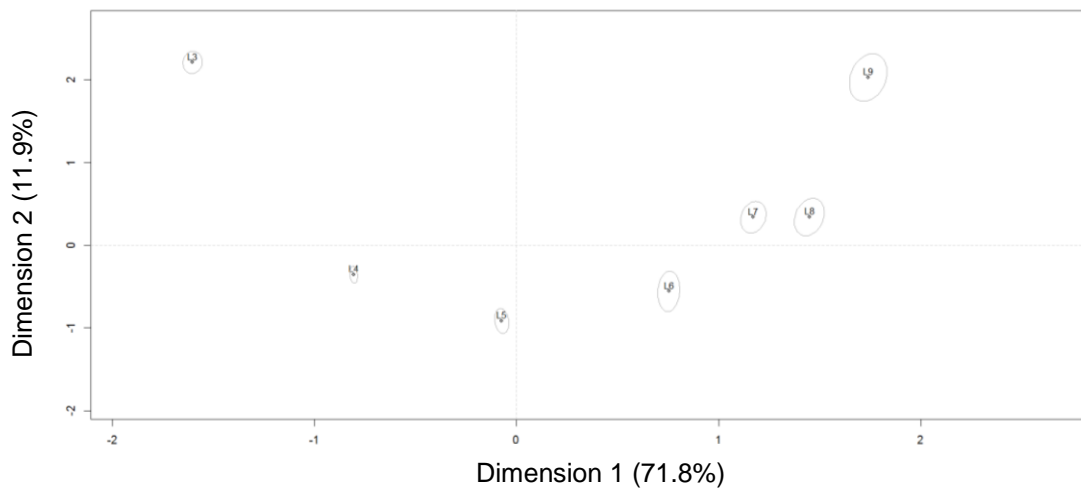


Figure 13. Results of the correspondence analysis (SST oral proficiency level).

Table 28. Coordinates for the SST Oral Proficiency Levels Ranked in the Score of Dimension 1

Level	Dimension 1	Dimension 2
3	-1.60	2.21
4	-0.81	-0.36
5	-0.07	-0.92
6	0.75	-0.56
7	1.17	0.34
8	1.45	0.34
9	1.74	2.03

The contribution rates show how much the results can explain the linguistic variation in the spoken texts. In other words, this is the extent that each factor contributes to explain the variance of the factors (Baayen, 2008). The more prominent dimension, Dimension 1, explained 71.8% of the linguistic variation and Dimension 2 accounted for 11.9%. The cumulative contribution rate of Dimension 1 and Dimension 2 totaled 83.7%. Thus, it successfully explains more than 80% of the variation in the cumulative contribution rate.

Next, Figure 13 shows that the oral proficiency groups (L3, L4, L5, L6, L7, L8, and L9) are distributed in order along the horizontal axis, Dimension 1. Smaller numbers, such as L3 and L4, which stand for lower level learners are positioned on the left side, and larger numbers, such as L7, L8, and L9, which stand for higher level learners are on the right side. The oral proficiency groups in the plot are distributed in a horseshoe-shape, not in a straight line. However, this phenomenon commonly occurs as a result of correspondence analysis (Camiz, 2005). There is debate over whether to accept this resulting phenomenon mathematically or not, but the resulting scatter plot in the present study does not show unusual data. The fact that they are in the right order shows that in general the use of grammatical features can distinguish oral proficiency levels, which is a reassuring finding for this study.

Another interesting finding is that most of the oral proficiency levels do not cluster in the plot, but L7 and L8 are displayed rather close to each other. The confidence intervals of L7 and L8, which are presented in balloons on the joint plot, do not overlap, so they are statistically distinct. However, this clustering of the L7 and L8 levels suggests

solely from a grammatical perspective, they are probably not as distinctive as the other levels in the Standard Speaking Test scale. Instead of seven levels from L3 to L9, there could be six levels - L9, L8/L7, L6, L5, L4, L3. This point is taken up again in the discussion chapter.

### **Correspondence Analysis Without Native Speakers of English**

It was confirmed that there is a general trend in which linguistic development over the different oral proficiency levels proceeds in a roughly left-right order of the linguistic features on the first dimension. Thus, by using the result of correspondence analysis, which does not include the frequency information of native speakers of English, I answer the first research question: what linguistic features characterize different English oral proficiency groups of Japanese learners? In order to answer this question I present the joint plot, the joint column and row plot that shows both. The results of this joint plot are zoomed to the center point (Figure 14). This joint plot can be used to find relationships between the oral proficiency levels and linguistic features. Table 29 shows the position of each linguistic feature in the plot, and the coordinate is ranked in the score of Dimension 1.

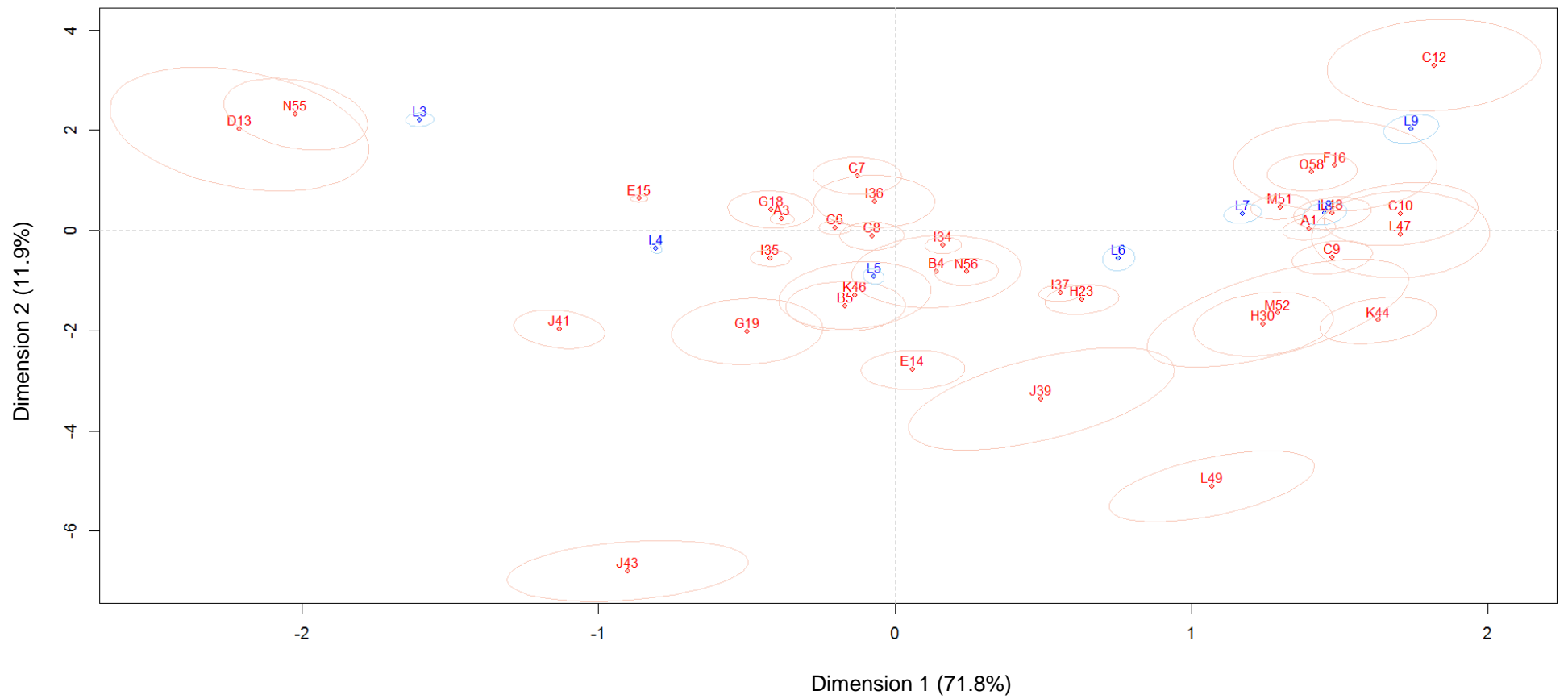


Figure 14. Results of the correspondence analysis (joint column and row plot).

Table 29. *Coordinates for the Linguistic Features Ranked in the Score of Dimension 1*

	Linguistic feature	Dimension 1	Dimension 2
H26	<i>that</i> relative in object position	11.34	32.55
H33	other adverbial subordinator	8.60	9.15
H27	WH relative in subject position	7.76	6.32
H20	<i>that</i> verb complement	7.76	6.32
H22	WH-clause	7.76	6.32
K45	necessity modal	7.29	2.86
H32	conditional adverbial subordinator	5.98	1.83
A2	perfect aspect	5.48	0.97
M54	split auxiliary	5.15	-0.81
O57	synthetic negation	4.91	-8.93
J38	Conjunct	3.57	-3.52
J42	Emphatic	3.20	1.12
J40	Hedge	3.12	-2.04
C11	indefinite pronoun	2.89	-1.58
C12	proverb <i>do</i>	1.82	3.30
C10	<i>demonstrative pronoun</i>	1.71	0.33
L47	public verb	1.70	-0.08
K44	possibility modal	1.63	-1.79
F16	agentless passive	1.48	1.29
C9	<i>pronoun it</i>	1.47	-0.54
L48	private verb	1.47	0.35
O58	analytic negation	1.41	1.16
A1	past tense	1.40	0.03
M51	Contraction	1.30	0.47
M52	stranded preposition	1.29	-1.65
H30	causative adverbial subordinator	1.24	-1.87
L49	suasive verb	1.07	-5.10
H23	infinitive ( <i>to</i> -clause)	0.63	-1.38
I37	Adverb	0.56	-1.24
J39	Downtoner	0.49	-3.36
N56	independent clause coordination	0.24	-0.82
I34	Preposition	0.16	-0.29
B4	place adverb	0.14	-0.82
E14	Nominalization	0.06	-2.78
I36	predicative adjective	-0.07	0.58
C8	third person pronoun	-0.08	-0.11
C7	second person pronoun	-0.13	1.09
K46	predictive modal	-0.13	-1.29
B5	time adverbial	-0.17	-1.51
C6	first person pronoun	-0.20	0.06
A3	present tense	-0.38	0.23
G18	<i>be</i> main verb	-0.42	0.42
I35	attributive adjective	-0.42	-0.55
G19	existential <i>there</i>	-0.50	-2.01
E15	Noun	-0.86	0.64
J43	discourse particle	-0.90	-6.79
J41	Amplifier	-1.13	-1.97
N55	phrasal coordination	-2.02	2.32
D13	direct WH-question	-2.21	2.02

The balloons on the joint plot of correspondence analysis show confidence intervals. Overlapping confidence intervals may indicate particularly strong

relationships between linguistic features and proficiency groups. L5 is associated with *predictive modal*, *place adverbial*, and *time adverbial*. These linguistic features are positioned near the center point of the joint plot, so it is suggested that they are used with much the same frequency by all groups (Yelland, 2010). They are all single word items that are not always necessary for constructing English sentences, but they are used to add extra information to the sentences.

Other than this finding, L7 is related to *contraction*. The box-and-whisker plot in Figure 15 shows overlap across the notches of the boxes from L7 to L9 learners, but the confidence intervals in Table 30 indicate statistically significant differences between L3/L4 (L3 upper: 1.95 and L4 lower: 2.07), L5/L6 (L5 upper: 2.41 and L6 lower: 2.61), and L6/L7 (L6 Upper: 2.9688 and L7 Lower: 2.9749).

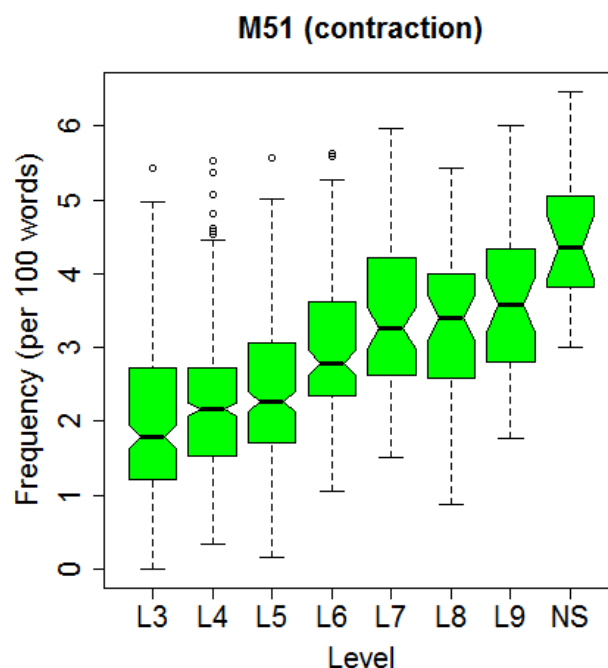


Figure 15. Box-and-whisker plots for M51 (contraction).

Consequently, it can be concluded that learners come to use *contraction* by the time they reach L7, and the learners persist in using it with similar frequency at higher levels. Expressed in a different way, L7 represents a turning point between the rise and the plateau for use of *contraction* on the box-and-whisker plots.

Table 30. *Confidence Intervals Across the Levels for M51 (Contraction)*

		L3	L4	L5	L6	L7	L8	L9	NS
M51	Upper	1.95	2.24	2.41	2.9688	3.55	3.70	3.96	4.79
	Lower	1.63	2.07	2.13	2.61	2.9749	3.10	3.20	3.91

Note. NS = Native speaker of English.

The close display of L7 and L8 in Figure 14 show that they are not as distinguished from each other as the other oral proficiency levels. However, L7 and L8 do not share the same linguistic feature. L8 was associated with following five items: *past tense*, *demonstrative pronoun*, *agentless passive*, *public verb*, and *private verb*, and one of them, *agentless passive*, is shared with L9. Among these linguistic items, in particular, *private verb* (e.g., *I assume that most of men, car is something like a precious for them, kind of like a accessory, or I don't know. 7\_0285.txt*) had a clear increasing frequency change pattern among the learners. Looking across the box-and-whisker plots in Figure 16, there are no overlaps across the notches of the boxes from L3 to L7 learners. The confidence intervals in Table 31 indicate statistically significant differences (L3 upper: 0.91 and L4 lower: 1.19, L4 upper: 1.32 and L5 lower: 1.35, L5 upper: 1.5313 and L6 lower: 1.5323, L6 upper: 1.84 and L7 lower: 2.00). Thus, it can be concluded that learners come to use L48 by the time they reach L7 (not L8 in this case), and then the use persists with similar frequency at higher levels.

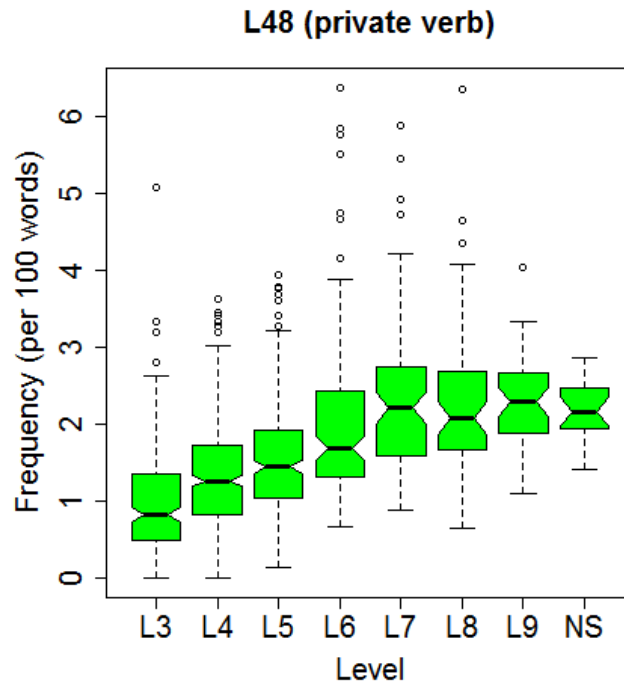


Figure 16. Box-and-whisker plots for L48 (private verb).

Table 31. Confidence Intervals Across the Levels for L48 (Private Verb)

		L3	L4	L5	L6	L7	L8	L9	NS
L48	Upper	0.91	1.32	1.53	1.84	2.41	2.29	2.47	2.34
	Lower	0.73	1.19	1.35	1.53	2.00	1.86	2.09	1.97

Note. NS = Native speakers of English.

Similar relationships between oral proficiency levels and some of the variables are displayed at points along and close to the Dimension 1 axis. Not only *contraction* and *private verb* but also some of the features mentioned in previous section as “weaker” relationships, such as *agentless passive*, *demonstrative pronoun*, and *past tense*. This is one of the great benefits of using correspondence analysis. It helps to see patterns in the box-and-whisker plots that may not be noticeable at first because of “slightly irregular” patterns in the box-and-whisker plots.

As pointed out in the literature review, it was found by the computer-aided error analysis which used the same NICT JLE corpus as this study that some linguistic categories (e.g., *voice*, *modal verb*, *finite verb* and *nonfinite verb*, *verbal form*, *verb complement*) can be more clearly understood when well-formed language use is focused on (Abe, 2007a). Taking into account the results of this analysis, *agentless passive* and *predictive modal* supported the findings of the previous study. It also supported the results of Negishi (2012), which suggested that *modal auxiliary* can effectively discriminate A2 to B2+ level learners of CEFR. According to the box-and-whisker plot analyses in the previous section, *noun*, *emphatic*, and *pronoun it* were considered to characterize different oral proficiency groups of Japanese learners. In addition to these findings, *contraction* and *private verb* can be added to this list.

### **Correspondence Analysis with Native Speakers of English as Supplementary Variable**

The second research question asked to what degree the language features appearing in the spoken production of high proficiency learners match with those of native speakers of English who perform the same task. Adding the frequency information of native speakers of English as a supplementary variable, another joint plot was created to check this issue. The results of the joint plot are zoomed to the center point (Figure 17). This second correspondence analysis was conducted to examine if the language features appearing in the high proficiency learners' data match with those of native speakers of English. However, the two most proficient groups, L8 and L9, are not linked with any of those linguistic features. Native speakers of English were related to *stranded preposition* (M52) (e.g., the candidate

that I was thinking *of*), and there was a statistical difference in frequency between L9 and native speakers of English. Thus, *stranded preposition* shows that the language features appearing in the spoken production of high proficiency learners do not match with those of native speakers of English. However, looking at the language features which are close to L9 and NS on the correspondence analysis plot, such as *demonstrative pronoun* (C10) and *proverb do* (C12), they are mostly used with about the same frequency by L9 learners and native speakers of English. In other words, the spoken production of the high proficiency learners with these language features appears to be similar to that of the native speakers of English.

As the results of correspondence analysis cannot clearly show the differences between native and non-native speakers of English, I list some of the linguistic features that show significant differences between them by checking the box-and-whisker plots. The plots show there are six linguistic features that native speakers of English use more frequently than L9 learners: *perfect aspect*, *place adverb*, *pronoun it*, *stranded preposition*, *synthetic negation*, and *emphatic*, and there are five that are used less frequently: *past tense*, *first person pronoun*, *infinitive*, *possibility modal*, and *analytic negation*. Many other linguistic features are used with similar frequency. To sum up, only 11 out of 49 linguistic features appearing in the spoken production of high proficiency learners do not match with those of native speakers of English who perform the same task, and three quarters of the analyzed linguistic features that L9 learners use match with those of native speakers of English. Accordingly, it can be suggested that the spoken production of the high proficiency learners appears to be similar to that of the native speakers of English.

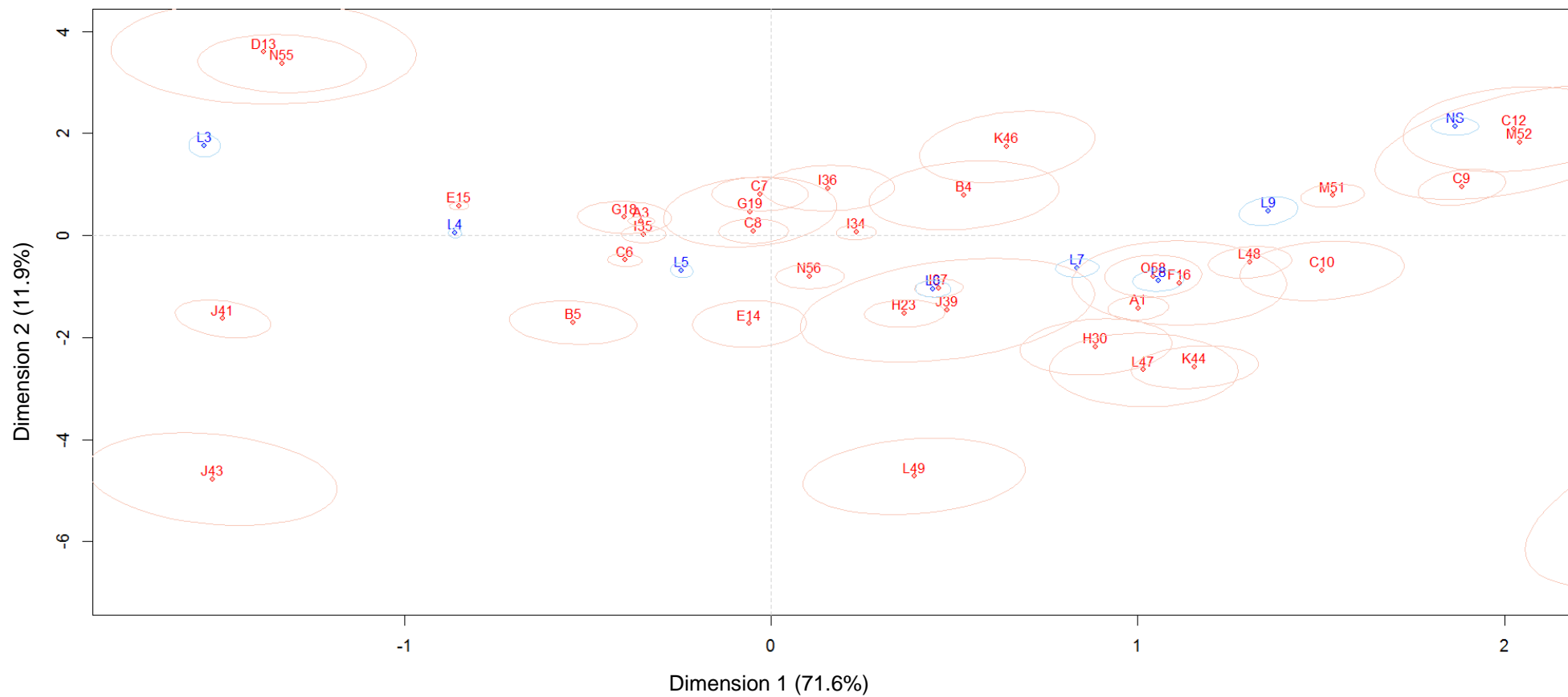


Figure 17. Results of the correspondence analysis (joint column and row plot with native speakers of English).

Native speakers of English used *first person pronoun* markedly differently from non-native speakers of English. As shown in Figure 18, *first person pronoun* does not show equal amounts of usage and it is not part of a falling or rising pattern where the native speakers' frequency of use is clearly different. The notches of the boxes indicate this significant difference.

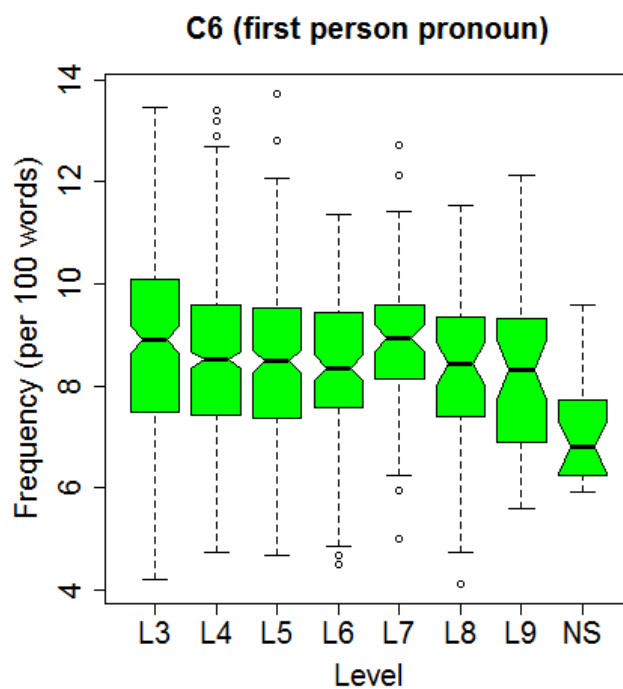


Figure 18. Box-and-whisker plots for C6 (first person pronoun).

Native speaker use of *first person pronoun* is much lower than non-native speaker use which is roughly level across the L3 group (e.g., “*I like winter. I like skiing or snowboarding. So in winter, I can skiing and I can snowboarding. So I like winter. I don’t like summer so much.*” 3\_0160.txt) to L9 groups (e.g., “*I mainly cook at home. And I’m organic, too. So I order organic foods and I’d rather not eat food outside because I know*

it's poisoned, as *I* would put it. Genetically modified products are out there. *I* don't trust it very much. So *I* order everything, and *I* would much rather have my kids eat the same as *I* do. So *I* enjoy cooking at home under the very busy circumstances. *I* try as hard as *I* can." 9\_1423.txt). This language use suggests that Japanese learners of English have a tendency to concentrate on using *first person pronoun* during their conversation. The highest level, L9 learners, can use impersonal pronouns, particularly *indefinite pronouns*, with the same frequency as native speakers of English to generalize their talk (e.g., *But everybody brings something.*). However, they still have a heavy dependency on using *first person pronoun*. Thus, *first person pronoun* can be considered as one of the features that distinguishes native and non-native speakers of English. Unlike this high dependency of L9 learners on *first person pronoun*, *pronoun it* is listed as an item that native speakers of English use more frequently than L9 learners, which involves the use of reference to connect to previous nouns, phrases or clauses. It was suggested in previous section that the falling frequency change pattern of *noun* is a sign of increasing proficiency. Thus, it can be assumed that learners come to replace *noun* by *pronoun it* as their oral proficiency increase, but still there remains a significant difference in the use of *pronoun it* between native and non-native speakers of English.

## CHAPTER 5

### DISCUSSION

Biber's (1988) list of linguistic features was used to find those that can characterize English oral proficiency levels for Japanese learners of English. This chapter contains an overview of the results of the study with answers to each of the research questions. Then, the strengths and weaknesses of correspondence analysis and box-and-whisker plot analysis are discussed with the aim of seeing how these two kinds of data analysis can complement each other.

#### Answers to the Research Questions

##### Research Question 1

The first research question is to identify the linguistic features that can characterize different English oral proficiency groups of Japanese learners. It was found that some linguistic features can clearly show the differences: *noun*, *emphatic*, and *pronoun it*. Among these features, *noun* presented a falling frequency change pattern and the others, *emphatic* and *pronoun it*, an increasing frequency change pattern.

*Noun* is one that lower oral proficiency learners used more frequently than higher oral proficiency learners and native speakers of English. It has been suggested that the frequency of different parts-of-speech can be used as a marker of interlanguage development in previous studies. It was suggested that control of the parts-of-speech develops in the order of nouns, verbs, and prepositional phrases in Tono (2000), which

focused on the written production of Japanese learners of English. Another study, which focused on the spoken production of Japanese learners of English, also suggested that learner language develops in the order of nouns to verbs (Kobayashi, 2007). In addition to these findings, Negishi (2012) suggested that two grammatical features, relative clauses and modal auxiliaries, out of 42 criterial features can effectively discriminate the A2 to B2+ levels of CEFR. The results of this study lend support to the idea that the frequency of some features could be used to distinguish oral proficiency levels. Nouns are more frequently used by lower level learners, which is partly because they struggle to complete some sentences with other parts of speech. They can start speaking with a subject (a *noun*), but find difficulty continuing. The proportional decrease of nouns in oral performance might be caused by a greater use of pronouns, coordinated predicates, and non-finite subjectless clauses at the higher oral proficiency levels. Thus, unlike *noun*, the frequency of *pronoun it* increases as the oral proficiency of learners rises.

Other than *noun* and *pronoun it*, *emphatic* is more frequently used by higher oral proficiency learners. The increasing use of *emphatic* is a sign of increasing oral proficiency, and it can be also considered as a distinctive feature for native speakers of English. However, it was difficult to identify the linguistic features that can characterize some of the oral proficiency groups, because the learners in the intermediate proficiency ranges did not differ much from each other, except by small amounts in the quantity of use of some features. That is, any differences in the occurrence of grammatical features are likely to be small differences among the middle proficiency groups in terms of relative frequency, rather than differences in occurrence versus non-occurrence of certain

features. According to the results of correspondence analysis, L7 and L8 are clustered so close to each other that they could be integrated into one level. This suggests solely from a grammatical perspective that there are probably too many proficiency levels in the Standard Speaking Test scale. Instead of seven levels from L3 to L9, there could be six levels – L9, L8/L7, L6, L5, L4, and L3.

However, reducing the number of levels on the speaking test is not completely desirable for two reasons. First, if the amount of study time necessary to improve from one level to another is roughly equivalent along the entire scale, then the elimination of proximate levels would result in having levels that were too far apart as regards the amount of study required to make recognizable progress. Second, the retention of several grammatically narrow levels could be beneficial to teachers and learners as they could consciously target grammatical elements associated with higher levels. The optimal number of levels should be based on the practical issue of accuracy of classification, but investigating this was beyond the aims of this study. The clustering of the mid-proficiency levels, which is solely based on a grammatical perspective, suggests a weakness in the scale. It would be interesting to see in a reliability study of test scoring if the most unstable part of the scale was the middle region of the scale.

A clustering of the levels suggests a weakness in the SST scale, but it could also suggest that the linguistic features investigated might not be the most appropriate ones for distinguishing between the proficiency groups. Thus, it is necessary to examine whether the linguistic features which have been shown to distinguish style and text variation of language in Biber (1988) and in previous research can also be used to distinguish oral

proficiency levels of Japanese EFL learners, which leads to the third research question. Also, there is another possibility that some other elements beyond the list have a much stronger effect on oral proficiency development and its assessment.

What is most striking in the normalized frequency data of the grammatical features across oral proficiency levels are the rather similar frequencies of occurrence.

Table 32 contains three lines from Table 28 to illustrate this point.

Table 32. *Normalized Frequency Counts in One Million Words of Three Grammatical Features with Similar Frequencies of Occurrence*

	Linguistic Features	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	NS
C7	second person pronoun	14,913	13,123	12,595	14,830	15,024	17,713	16,380	15,087
C8	third person pronoun	22,695	23,799	24,558	24,312	22,519	23,053	26,474	23,592
C9	pronoun <i>it</i>	9,428	12,403	14,815	16,208	19,779	19,595	21,399	26,966

Note. NS = Native speaker of English.

*Second person pronoun* and *third person pronoun* are not distinctive features and their frequencies are similar across the levels. *Pronoun it* is a distinctive feature for higher proficiency learners and native speakers of English, but even the L3 learners use the *pronoun it* with a reasonable frequency of occurrence. Occurrence versus non-occurrence is not a characteristic of the data when I look at the oral proficiency groups. Language use involves a variety of grammatical constructions and more complex grammatical features do not replace simpler ones but complement them. This result may arise because some more complicated features might be used by lower proficiency learners without full control of their use or in very limited circumstances. For example they could be used in formulaic sequences that are correct but largely unanalyzed.

A striking feature of many of the box-and-whisker plots is the narrow range of scores of the native-speaker group even when their median is similar to the other groups. For example, the *third person pronoun* plot shows the narrow range of scores of native speakers. This is due in part to the small number of people in the native speaker group, but the greatest spread of scores (the biggest boxes and longest whiskers) is typically in the lowest proficiency group (L3) and this was not the largest group but the third in group size (the first in group size is L4: 482 learners, the second is L5:236 learners, and the third is L3: 222 learners). This shows that although the *n*-size of the groups might have been a factor in the range of scores within a group, oral proficiency level was also an important factor. Unsurprisingly, the learners in the lowest proficiency group show greater variability in having gained control of various linguistic features.

The rating of the SST was done holistically using the criteria presented in Appendix A. Although these criteria are generally not described using linguistic features, various linguistic features that can characterize the oral proficiency levels were found in the present study. They can be grouped into the following categories: (a) part-of-speech (*noun, pronoun it, first person pronoun, demonstrative pronoun, indefinite pronoun, possibility modal, adverb, causative adverb*), (b) stance marker (*emphatic, hedge, amplifier*), (c) reduced forms (*contraction, stranded preposition*), (d) specialized verb classes (*private verb*), complementation (*infinitive*), (e) coordination (*phrasal coordination*), (f) passive (*agentless passive*), and (g) possibly tense and aspect markers (*past tense, perfect aspect*). The SST assessment criteria used general terms (e.g., *mostly correct simple sentences; good command in using tenses; omission of conjunctions*;

*elementary errors sometimes occur in complex sentences but not habitual; can produce grammatically correct speech unconsciously*) to describe both well-formed and erroneous language use. The results of the present study are clearly not the effect of finding grammatical features that are explicitly contained in the rating criteria. The criteria are not expressed in that way.

The use of each feature at the various proficiency levels was analyzed to see if there were patterns across the levels. In addition to a flat pattern, and simple falling and rising frequency patterns, other kinds of patterns were found: (a) rise-plateau (*infinitive, possibility modal, agentless passive, adverb, causative adverbial subordinator, hedge, contraction, indefinite pronoun, demonstrative pronoun, private verb, and possibly past tense*), (b) plateau-fall (*amplifier*), (c) fall-plateau (*phrasal coordination, present tense*), and (d) plateau-rise (possibly *perfect aspect*). It was found that not only the falling (*noun*) and rising (*pronoun it, emphatic*) frequency change pattern but also the combination of rise, fall, and plateau patterns can show the differences in learner language use. These combination patterns can be included in the group of falling and rising frequency change patterns.

The answer to the first research question is that it is possible to identify the linguistic features that can characterize different English oral proficiency groups of Japanese learners.

## Research Question 2

The second research question is to find to what degree the language features appearing in the spoken production of high proficiency learners match with those of native speakers of English. There are some grammatical features that distinguish the high proficiency learners and native speakers of English. According to the results of correspondence analysis, native speakers of English and L9 learners were positioned on the same side of the vertical axis with the L9 learners much closer to the centre point, but correspondence analysis did not provide so much useful information on answering the second research question. Accordingly, the difference of L9 and native speakers of English was presented by the confidence intervals. The following features were able to distinguish the differences: (a) linguistic features that native speakers use more frequently than L9 learners (*perfect aspect, place adverb, pronoun it, stranded preposition, synthetic negation, emphatic*) and (b) linguistic features that native speakers of English use less frequently than L9 learners (*past tense, first person pronoun, infinitive, possibility modal, analytic negation*). In addition to these features, many other linguistic features were used with similar frequency. Thus, it can be suggested that the spoken production of high proficiency learners with these language features appears to be similar to that of the native speakers of English.

The box-and-whisker plots of native speakers of English sometimes show a more homogeneous use when they are compared with those of Japanese learners of English. For example, the box-and-whisker plots of *demonstrative pronoun* (Figure 9) are compact and native speaker use is rather similar (and low), while there is clearly great variety in

the use of *demonstrative pronoun* in language learner groups (e.g., the length of whiskers). Additionally, the increases in overall proficiency may sometimes be accompanied by increasing differences from the norms of native speakers of English.

The answer to the second research question is that the language features appearing in the spoken production of high proficiency learners match with those of native speakers of English to a small degree.

### **Research Question 3**

The third research question is to know if the oral production of Japanese EFL learners is rich enough to display the full range of features used by Biber. With one exception, all of Biber's 58 features occurred in each of the oral proficiency levels. The only exception was *WH relative with fronted preposition*, a low-frequency item which did not occur at the L3 level (Table 25). However, it was not worth keeping all the 58 linguistic features in the analysis, because they did not all occur with sufficient frequency to make a sensible comparison across the proficiency levels. Nine of the linguistic features were deleted from the analysis because their medians in the learner groups were zero and they did not discriminate between the oral proficiency groups. There are many examples of particular learners not using a grammatical feature. This of course can come from lack of knowledge of the item, and lack of opportunity to use the item either because of the content of the topic or because the item is typically used with very low frequency by even accomplished native speakers of English. Consequently, it can be

concluded that the oral production of Japanese EFL learners is rich enough to display the full range of features used by Biber.

Learners at the various proficiency levels used the linguistic features in Biber's list. The frequency of the 49 linguistic features was high enough to conduct a corpus-based learner language analysis. What is clear is that the tasks used in the speaking test and the time allowed for the tasks provided ample opportunity for Biber's features to be used across the oral proficiency levels, and thus the list is useful for studying the variation of oral performance produced by Japanese learners of English.

It is worth noting in the case of learner language, linguistic features that characterize different processing modes are often seen in both written and spoken production. In a previous study, Abe (2008) investigated learner corpora consisting of a written composition and oral production drawn from a picture description task completed by the same Japanese EFL learners. She found that the linguistic features of both careful and vernacular styles were seen in both the spoken performance and written composition of Japanese EFL learners. The concept of an interlanguage continuum needs to incorporate the idea that both careful (i.e. written) and vernacular (i.e., spoken) styles make up the interlanguage system (Tarone, 1983). Accordingly, in the case of learner language, it is reasonable to use Biber's list, because it contains linguistic features that can characterize both spoken and written processing modes.

The answer to the third research question is that the oral production of Japanese EFL learners as produced in the speaking test is rich enough to display the full range of features used by Biber.

## **The Strengths and Weaknesses of Correspondence Analysis and Box-and-whisker Plot Analysis**

The aim of the present study was to find linguistic features that can characterize a range of English oral proficiency levels and to typify oral proficiency levels using a large number of grammatical features from a spoken corpus containing over one-million-tokens. Thus, it was necessary to choose a statistical technique which is effective in handling a large number of data points that are not easily analyzed through a visual inspection or a series of pairwise comparisons of variables. Additionally, it was necessary to choose a method that is useful for examining the interrelationships among English oral proficiency levels and the multiple linguistic features focused on in the study. Correspondence analysis has an advantage in graphically showing similarities and dissimilarities among variables on a plot, and correspondence analysis makes it easier to see which items distinguish each group, and how the groups relate to each other.

However, correspondence analysis has weaknesses. These are largely because correspondence analysis has to reach a compromise between numerous data points in order to be able to display the results in a multiple dimensional plot. In addition, in order to compare two variables, item frequency and proficiency level as was the case in this study, individual results had to be pooled into oral proficiency groups. In the interests of gaining a broad picture, detail is lost. Correspondence analysis is effective in signalling distinctiveness and providing a general picture of multivariate data. Correspondence analysis gave a general picture of how language use shifted across the oral proficiency

groups. Thus, it was essential to use another statistical technique to add detail to the picture it gave.

Box-and-whisker plots can provide a detailed examination of each grammatical feature (e.g., medians, the size of the boxplots, the length of the whiskers, the number of outliers, the spread of use within a proficiency level, and whether there were no or very low numbers of occurrences). What is more, frequency patterns (e.g., rise, flat, and fall) across the levels can be indicated by a set of box-and-whisker plots, and the relationship between the notches of the boxes can indicate whether there is a significant difference between them.

The data for correspondence analysis come from a data matrix showing the frequency for each grammatical item at each oral proficiency level. This is useful to go back to when considering the various distinctive features because these frequency figures show whether the item is a low frequency item or not and whether there were zero occurrences for any items at any levels.

These three ways of examining the data, correspondence analysis, box-and-whisker plots and the data matrix, usefully complement each other. They differ in that the data matrix provides raw (or adjusted) frequency figures, correspondence analysis is an exploratory approach that provides questions that can be followed up by other statistical methods, and box-and-whisker plot analysis allows a detailed examination of each linguistic feature. These approaches thus nicely complement each other. They provide different viewpoints of essentially the same information.

## CHAPTER 6

### CONCLUSION

#### Summary of the Findings

There were three main findings. First, some interesting frequency change patterns (i.e., a rising, a falling, a combination of rising, falling, and plateauing) across the oral proficiency level were shown by the following linguistic features: *noun, emphatic, pronoun it, infinitive, possibility modal, agentless passive, adverb, causative adverbial subordinator, hedge, contraction, indefinite pronoun, demonstrative pronoun, private verb, amplifier, phrasal coordination, present tense, possibly past tense and perfect aspect*. Even though there is a difference in terms of strength in characterizing English oral proficiency groups of Japanese learners, these linguistic items are revealing. The second finding concerned the gap between native and non-native speakers of English. There are six linguistic features that native speakers use more frequently than L9 learners (*perfect aspect, place adverb, pronoun it, stranded preposition, synthetic negation, emphatic*) and there are five linguistic features that native speakers of English use less frequently than L9 learners (*past tense, first person pronoun, infinitive, possibility modal, analytic negation*). However, many of the other linguistic features were used with similar frequency by L9 learners and native speakers of English suggesting that in several ways the spoken production of high proficiency learners appears to be similar to that of the native speakers of English. Third, it was found that the oral production of Japanese EFL learners is rich enough to display the full range of features used by Biber.

The results of this study suggested the use of different linguistic features can distinguish the different oral proficiency levels and native and non-native speakers of English. For example, the most frequently used feature, *noun*, was more frequently used by lower proficiency learners than higher proficiency learners and native speakers of English. The frequency decreased as proficiency level rose. In contrast, *emphatic* and *pronoun it* were more frequently used by native speakers of English, and their frequency increased as oral proficiency rose. The increases in overall proficiency may sometimes be accompanied by increasing differences from the norms of native speakers of English as in these items. To sum up, various types of frequency change patterns of linguistic features that can reflect the variation of learner language were found in the present study.

### **Limitations**

In terms of linguistic features, the present study did not distinguish lexical, syntactic, and morphological effects. There is also a lack of detail in some of the syntactic categories, for example, regular and irregular past verb forms were not divided in the past tense category.

Another feature that was not included in the present study is learner errors. Finding that a particular form is frequently used at a particular proficiency level does not always mean that the form has been used correctly. Researchers can gain reasonably accurate frequency information on each linguistic feature, but as is the case with the past form for example, it is necessary to read the whole context and compare several past tense forms that are possibly not grammatically correct may be counted to see if they are

used accurately. Fortunately, this study is complemented by Abe's (2007a) study, which examined the same NICT JLE corpus by manually annotating 31 different types of errors. It was found that particular error types can be markers of a specific English oral proficiency level, and learner language is likely to be much better understood when it is focused on well-formed language use as well as on ill-formed language use. Now that we have a list of features distinguishing proficiency levels, the accuracy of use of these features can be a focus for future research.

This study was based on an oral proficiency test that contains various elicitation tasks. The oral proficiency test did not always use the same elicitation tasks, and this can cause unnecessary variations in the performance of test-takers. Additionally, a particular elicitation task might emphasize particular vocabulary or grammatical structures and de-emphasize the others. There is also a possibility that some of the participants might have spent more time on description rather than narration during the story-telling task. Thus, the different oral proficiency levels might be distinguished by a different focus on various aspects of the task. In other words, speakers at different proficiency levels might have focused on different "genres" within the task. Consequently, there is a possibility that the speaking task can have some influence on the results.

Another limitation is that the present study did not distinguish between memorized routines and grammatical usage. For instance, *what* might often be found in the phrase "*What is ...*," and this would make *what* highly frequent, but would also likely make it a memorized routine. It is said that formulaic speech plays a major role in language learning, and it can facilitate the linguistic competence of language learners.

Thus, it is important to distinguish memorized routines from creative usage. However, as pointed out in Ellis and Barkhuizen (2005), it is very difficult to distinguish ‘creative’ and ‘formulaic’ speech in learner performance. Formulaic speech can evolve into creative usage in the process of language learning. I made no attempt to differentiate these two types of production.

As a final limitation, the normative data were obtained from only 20 native speakers who performed the same speaking tasks as the learners. From oral proficiency level 4 onwards there was a drop in the number of learners, and this is likely to contribute at least in a small way to a narrowing of the range of scores as proficiency increases and *n*-size decreases. The average length of the spoken text increased as oral proficiency increased, but it is still necessary to bear in mind the number of learners and text size at each level when interpreting the analysis.

Box-and-whisker plots and correspondence analysis were used to identify the linguistic features that can characterize the variation of oral proficiency groups. These two statistical analyses differ in that the box-and-whisker plots allow a detailed examination of each linguistic feature and correspondence analysis provides an overall picture of linguistic feature. However, oral proficiency levels of all test-takers were provided in this study. The choice of any statistical technique brings advantages and limitations. By using two statistical techniques, box-and-whisker plots and correspondence analysis, the limitations are reduced with one compensating for the other. In a future study, regression analysis could be used to examine the relationship of the proficiency groups and the linguistic features.

### **Suggestions for Future Research**

Laufer and Nation (1995) advocated an index, the Lexical Frequency Profile (LFP), for measuring the lexical richness of L2 written production. McCarthy and Jarvis (2007) evaluated the reliability of the D-Index to measure lexical diversity to solve the type-token ratio (TTR) problem caused by different text lengths. However, relatively little research has analyzed interlanguage development from the aspects of both lexis and grammar. Accordingly, it is worth examining lexical and lexico-grammatical patterns. They can be compared based on the frequently used words, part-of-speech bigrams and trigrams, lexical frequency profiles, and this additional information might help distinguish oral proficiency groups.

Additionally, considering naturalness in spoken English, it is useful to understand which collocations are much more frequently used across the oral proficiency levels. Also, some linguistic features, such as articles and demonstrative determiners, which are considered to be difficult for Japanese EFL learners to acquire, should be analyzed in future studies. What is more, I did not examine how the task differences influenced the language features that test-takers use. Therefore, how each linguistic feature is distributed across different speaking tasks and different oral proficiency levels should be examined in future. Finally, it is worth explaining how learner language shifts across the oral proficiency levels from the perspectives of input language, cognitive factors, and the influence of the first language.

## **Final Conclusions**

In order to fill a gap revealed in an examination of previous studies, I used language processing techniques to extract the frequency information of 58 linguistic features used by EFL learners and native speakers of English from a one-million word spoken corpus. The linguistic features were those used by Biber (1988) to describe variation in a wide range of spoken and written texts produced by native speakers. It was found that there are grammatical features that can be used to distinguish a wide range of oral proficiency levels of Japanese EFL learners. The use of most of these distinguishing features can be explained, at least retrospectively. Grammatical features alone are not enough to fully distinguish oral proficiency levels, yet the results of this study show that certain grammatical features can be useful oral proficiency level signals.

Understanding how learner language differs from the lower level to the advanced level is important because this understanding can contribute to more effective language learning and language teaching. During the process of language learning, it is useful for learners to understand the norms of the target language and the characteristic differences between the interlanguage and the target language. It is also necessary for researchers to understand these differences to develop appropriate language assessment materials.

It was also shown that a methodological approach combining electronic learner language data, language processing techniques, multivariate statistical analysis, and visual inspection of descriptive statistics is useful in exploring learner language variation and the differences between native and non-native speakers of English. In other words, it showed how a learner corpus can be used to reveal learner language variation by

describing the characteristics of interlanguage. It is hoped that the results of this study can contribute to an understanding of the nature and characteristics of learner language variation across a wide range of linguistic features.

## REFERENCES

- Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 132-141). London: Addison Wesley Longman.
- Abe, M. (2007a). A corpus-based investigation of errors across proficiency levels in L2 spoken production. *JACET Journal*, 44, 1-14.
- Abe, M. (2007b). Grammatical errors across proficiency levels in L2 spoken and written English. *The Economic Journal of Takasaki City University of Economics*, 49, 117-129.
- Abe, M. (2007c). JEFLL コーパスに見る品詞別エラーの全体像 [An overview of part-of-speech error in JEFLL corpus]. In Y. Tono (Ed.), 日本人中高生一万人の英語コーパス “JEFLL Corpus”—中高生が書く英文の実態とその分析— [JEFLL corpus, a corpus of 10,000 Japanese junior and senior high school students: Analyzing written composition of junior and senior high school student] (pp. 146-158). Tokyo: Shogakkan.
- Abe, M. (2008, March). *An analysis of linguistic features in identical picture description tasks: The oral and written production of non-native and native speakers of English*. Paper presented at the meeting of the PacSLRF2008, the 3rd National Symposium on SLA, Beijing.
- Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55-76). Amsterdam: Benjamins.
- Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring large educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, 1(1), 96-129. doi:10.1075/ijlcr.1.1.04ale
- Altenberg, B. (1989). Review of D. Biber (1988), Variation across speech and writing. *Studia Linguistica*, 43(2), 167-174. doi:10.1111/j.1467-9582.1989.tb00800.x
- Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 37-53). Amsterdam: Benjamins.

- Asención-Delaney, Y. (2015). A multi-dimensional analysis of advanced written L2 Spanish. In T. B. Sardinha, & M. V. Pinto (Eds.), *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber* (pp. 239-269). Amsterdam: Benjamins.
- Asención-Delaney, Y., & Collentine, J. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32(3), 299-322.  
doi:10.1093/applin/amq053
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, England: Cambridge University Press.
- Bendixen, M. (2003). A practical guide to the use of correspondence analysis in marketing research. *Marketing Bulletin*, 14, Technical Note 2.
- Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Lingua*, 62(2), 384-414.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Biber, D., & Conrad, S. (2009). *Register, genre, style*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.
- Biber, D., & Reppen, R. (1998). Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (Ed.), *Learner English on computer* (pp. 145-158). London: Addison Wesley Longman.
- Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33, 1-17. doi:10.1111/j.1467-1770
- Callies, M. & Götz, S. (Eds.). (2015). *Learner Corpora in Language Testing and Assessment*. Amsterdam: Benjamins.
- Camiz (2005). The Guttman effect: Its interpretation and a new redressing method. *Tetradia Analushs Dedomenwn [Data Analysis Bulletin]*, 5, 7-34.

- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English*. Cambridge: Cambridge University Press.
- Connor-Linton, J., & Shohamy, E. (2001). Register variation, oral proficiency sampling, and the promise of multi-dimensional analysis. In D. Biber & S. Conrad (Eds.), *Variation in English: Multi-dimensional studies* (pp. 124-137). Harlow: Pearson Education.
- Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics*, 5, 161-169.
- Cowie, A. P. (Ed.). (1998). *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Boston: Duxbury Press.
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010a). The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3), 573-605. doi:10.1111/j.1467-9922.2010.00568.x
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010b). The development of semantic relations in second language speakers: A case for Latent Semantic Analysis. *Vigo International Journal of Applied Linguistics*, 7, 55-74.
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2), 243-263. doi:10.1177/0265532211419331
- Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System: An International Journal of Educational Technology and Applied Linguistics*, 26, 163-174. doi:10.1016/S0346-251X(98)00001-3
- Daller, H, van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24(2), 197-222. doi:10.1093/applin/24.2.197
- De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). London, UK: Addison Wesley Longman.

- De Mönink, M., Brom, N., & Oostdijk, N. (2003). Using the MF/MD method for automatic text classification. In S. Granger & S. Petch-tyson (Eds.), *Extending the scope of corpus-based research: New applications, new challenges* (pp. 15-25). Amsterdam: Rodopi.
- Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards in the future. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 9-29). Amsterdam: Benjamins.
- Díaz-Negrillo, A. & Ballier, N., & Thompson, P. (Eds.). (2013). *Automatic treatment and analysis of learner corpus data*. Amsterdam: Benjamins.
- Dulay, H., & Burt, M. (1973). Should we teach children syntax? *Language Learning*, 23, 245-258. doi:10.1111/j.1467-1770.1973.tb00659.x
- ETS (2005). TOEFL Internet-based test: Score comparison tables. Retrieved July18, 2009 from ETS organization Web site:  
[http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL\\_iBT\\_Score\\_Comparison\\_Tables.pdf](http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Comparison_Tables.pdf)
- Ellis, R. (2008). *The study of second language acquisition*. Oxford: Oxford University Press.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.
- Garside, R., Leech, G., & McEnery, A. (Eds.). (1997). *Corpus annotation: Linguistic information from computer text corpora*. Harlow, England: Addison Wesley Longman.
- Ghadessy, M. (2003). Comments on Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt's "Speaking and Writing in the University: A Multidimensional Comparison." A Reader Reacts. *TESOL Quarterly*, 37(1), 147-150. doi:10.2307/3588469
- Gilquin, G., Papp, S., & Díez-Bedmar, M. (2008). Introduction. In G. Gilquin, S. Papp, & M. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. vii-xi). Amsterdam: Rodopi.
- Götz, S. (2015). Tense and aspect errors in spoken learner English: Implications for language testing and assessment. In M. Callies & S. Götz. (Eds.), *Learner corpora in language testing and assessment* (pp. 191-216). Amsterdam: Benjamins.

- Granger, S. (1994). The learner corpus: A revolution in applied linguistics. *English Today*, 10(3), 25-29.
- Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to bilingual and learner computerized corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press.
- Granger, S. (Ed.). (1998a). *Learner English on computer*. London: Addison Wesley Longman.
- Granger, S. (1998b). The computerized learner corpus: a versatile new source of data for SLA research, In S. Granger (Ed.), *Learner English on Computer* (pp. 3-18). London: Addison Wesley Longman.
- Granger, S. (1998c). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145-160). Oxford: Oxford University Press.
- Granger, S. (1999). Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In Hasselgård, H., & Oksefjell, S. (Eds.), *Out of Corpora: Linguistic information from computer text corpora* (pp. 191-202). Amsterdam: Rodopi.
- Granger, S., Dagneaux, E., & Meunier, F. (2002). The international corpus of learner English: Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). International corpus of learner English. (2nd version). Louvain-la-Neuve: Presses Universitaires de Louvain.
- Granger, S., Gilquin, G. & Meunier, F. (Eds.). (2013). Twenty years of learner corpus research - looking back, moving ahead. *Proceedings of the first learner corpus research conference (LCR 2011)*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: Benjamins.
- Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards

reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, 19(3), pp. 252-268. doi:10.1017/S0958344007000237.

- Granger, S., & Rayson, P. (1998c). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on computer* (pp. 119-131). London: Addison Wesley Longman.
- Gries, S. Th. (2008). Corpus-based methods in analyses of second language acquisition data. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 406-431). New York, NY: Routledge.
- Gries, St. Th. (2009). *Quantitative corpus linguistics with R: A practical introduction*. London & New York: Routledge, Taylor & Francis Group.
- Gries, S. Th., & Berez, A. L. (forthcoming). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation*. New York, NY: Springer.
- Hammond, M. (2003). *Programming for linguists: Perl for language researchers*. Oxford: Blackwell.
- Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 143-173). Amsterdam: Benjamins.
- Hawkins, J. A., & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English profile programme. In L. Taylor & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 158-175). New York, NY: Cambridge University Press.
- Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(1), 1-23.  
doi:10.1017/S2041536210000103
- Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the common European framework*. Cambridge: Cambridge University Press.
- Haiyang, A. & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students writing. In N. Ballier, A. Díaz-Negrillo, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 249-264). Amsterdam: Benjamins.

- Howarth, P. (1998). The phraseology of learners' academic writing and second language proficiency. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161-186). Oxford: Oxford University Press.
- Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 77-116). Amsterdam: Benjamins.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Husson, F., Lê, S & Pagès, J. (2011). *Exploratory Multivariate Analysis by Example Using R*. Boca Raton, FL: CRC Press.
- Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning and research* (pp. 3-11). Glasgow: University of Strathclyde Press.
- Ishikawa, S. (Ed.). (2013). *Learner corpus studies in Asia and the world, Vol.1*. Kobe: School of Language and Communication, Kobe University.
- Izumi, E., & Isahara, H. (2004a). Investigation into language learners' acquisition order based on an error analysis of a learner corpus. *Proceedings of the Pacific-Asia Conference on Language, Information and Computation, 18*.
- Izumi, E., Uchimoto, K., & Isahara, H. (2004b). 日本人1200人の英語スピーキングコーパス [A speaking corpus of 1200 Japanese learners of English]. Tokyo, Japan: ALC Press.
- Jarvis, S. (2002). Short text, best-fitting curves and new measures of lexical density. *Language Testing, 19*(1), 57-84. doi: 10.1191/0265532202lt220oa
- Kaneko, T. (2004). The use of past tense forms by Japanese learners of English. In Nakamura, J., Inoue, N., & Tabata, T. (Eds.), *English corpora under Japanese eyes* (pp. 215-230). Amsterdam: Rodopi.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. Harlow: Addison Wesley Longman.
- Kobayashi, Y. (2007a). The NICT JLE corpus と語彙研究 [The NICT JLE corpus and lexical study: A reexamination of the SST level]. 『英文学誌』 49, (pp. 17-29).

- Kobayashi, Y. (2007b). The NICT JLE Corpus における発達指標の研究—コレスポ  
 デンス分析によるタグ頻度解析 [Investigating developmental criteria in the  
 NICT JLE corpus through correspondence analysis of word-class distribution].  
 『言語処理学会第 13 回年次大会発表論文集』 [Annual report of the  
 association for natural language processing]. (pp. 486-489).
- Kobayashi, Y. (2010). コレスポデンス分析：データ間の構造を整理する [The  
 correspondence analysis: Summarize the structure among the data]. In S. Ishikawa,  
 T. Maeda, & M. Yamazaki (Eds.), 言語研究のための統計入門 [An  
 introduction to statistics for linguistic research] (pp. 245-264). Tokyo, Japan:  
 Kuroshio Shuppan.
- Lakshmanan, U., & Selinker, L. (2001). Analysing interlanguage: How do we know what  
 learners know? *Second Language Research*, 17, 393-420.  
 doi:10.1177/026765830101700406
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written  
 production. *Applied Linguistics*, 16, 307-322.
- Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A  
 corpus analysis of learners' English. *Language Learning*, 61(2), 647-672.  
 doi:10.1093/applin/16.3.307
- Lee, D. Y. W. (2003). *Modelling variation in spoken and written English*. Abington,  
 England: Routledge.
- Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer*. London:  
 Addison Wesley Longman.
- Lorenz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of  
 adjective intensification. In S. Granger (Ed.), *Learner English on computer* (pp.  
 53-66). London: Addison Wesley Longman.
- Lu, X. (2010). What can corpus software reveal about language development? In A.  
 O'Keefe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*  
 (pp. 184-193). Abingdon: Routledge.
- Lüdeling, A., Kyto, M., & McEnery, A. (forthcoming). (Eds.), *Handbooks of linguistics  
 and communication science volume corpus linguistics*. Berlin: Mouton de  
 Gruyter.

- Marsden, E. & David, A. (2008). Vocabulary use during conversation: A cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *Language Learning Journal*, 3(2), 181-98. doi:10.1080/09571730802390031
- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. 3999*Language Testing*, 24, 259-488.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Oxford: Routledge.
- Meurers, D. (2009). On the Automatic Analysis of Learner Language: Introduction to the special issue. *CALICO Journal*, 26(3), 469-473.
- Meurers, D. (2013). Natural language processing and language learning. In C. A. Chapelle (Ed.), *Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell.
- Meurers, D. (forthcoming). Learner Corpora and Natural Language Processing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.
- Meunier, F., & Littre, D. (2013). Tracking learners' progress: Adopting a dual 'corpus cum experimental data' approach. *The Modern Language Journal*, 97, 61-76. doi:10.1111/j.1540-4781.2012.01424.x
- Mizumoto, A. (2009). コーパス言語学研究における多変量解析手法の比較—主成分分析 vs. コレスポネンス分析 [Comparison of multivariate data analysis methods in corpus linguistics: Principal component analysis vs. correspondence analysis]. 『コーパス言語研究における量的データ処理のための統計手法の概観』 [An overview of statistical methods for corpus linguistics] *The institute of statistical mathematics cooperative research report 232* (p. 53-64). Tokyo, Japan: The Institute of Statistical Mathematics.
- Murakami, A. (2009). A corpus-based study of English textbooks in Japan and Asian countries: Multidimensional approach. (Unpublished master's thesis). Tokyo University of Foreign Studies, Tokyo.
- Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, 21(4), 373-391. doi:10.1191/0267658305sr252oa

- Nakamura, J. (1995). Text typology and corpus: A critical review of Biber's methodology. *English Corpus Studies*, 2, 75-90.
- Nakamura, J. (2002). A galaxy of words: Structures based upon distributions of verbs, nouns and adjectives in the LOB corpus. In T. Saito, J. Nakamura & S. Yamazaki (Eds.), *English corpus linguistics in Japan* (pp. 19-42). Amsterdam: Rodopi.
- Nakamura, J. (2005). 検索したデータを分析する [Analyzing retrieved data.] In T. Saito, J. Nakamura & I. Akano (Eds.). (2005). 『英語コーパス言語学—基礎と実践』 [*English corpus linguistics: Basic and practice* ] (pp. 92-117). Tokyo, Japan: Kenkyusha.
- Neselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: Benjamins.
- Negishi, M. (2012). CEFR 基準特性に基づくチェックリスト方式による英作文の採点可能性 [Exploring the possibility of assessing English writing based on the checklist of CEFR criterial features.] *ARCLE REVIEW*, 6, 90-99.
- Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA. *Applied Linguistics*, 30, 555-578. doi:10.1093/applin/amp044
- Pendar, N., & Chapelle, C. (2008). Investigating the promise of learner corpora: Methodological issues. *CALICO Journal*, 25(2), 189-206. doi:10.1558/cj.v25i2.189-206
- Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Harlow: Pearson Education.
- Renouf, A., & Kehoe, A. (Eds.). (2009). *Corpus linguistics: Refinements and reassessments*. Amsterdam: Routledge.
- Ringbom, H. (1998). Vocabulary frequencies in advanced learners English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41-52). London: Addison Wesley Longman.
- Saito, T., Nakamura, J., & Akano, I. (Eds.). (2005). 『英語コーパス言語学—基礎と実践』 [*English corpus linguistics: Basic and practice* ]. Tokyo, Japan: Kenkyusha.
- Saito, T., Nakamura, J., & Yamazaki, S. (Eds.). (2002). *English corpus linguistics in Japan*. Amsterdam: Rodopi.

- Salamoura, A., & Saville, N. (2009). Criterial features of English across the CEFR levels: Evidence from the English Profile Programme. *Research Notes*, 37, 34-40.
- Salamoura, A. & Saville, N. (2010). Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. In I. Bartning, M. Martin, & I. Vedder. (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 101-132). Eurosla monographs series.
- Sardinha, T. B., & Pinto, M. V. (Eds.) (2014). *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*. Amsterdam: Benjamins.
- Santorini, B. (1991). *Part-of-speech tagging guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.
- Schmid, H. (1994). *Probabilistic part-of-speech tagging using decision trees. Revised version of a paper presented at the International Conference on New Methods in Language Processing*, Manchester, England.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10, 209-232. doi:10.1515/iral.1972.10.1-4.209
- Sugiura, M. (Ed.). (2008). 英語学習者のコロケーション知識に関する基礎的研究 [Basic research on collocation knowledge of L2 English learners] Nagoya: Nagoya University.
- Svartvik, J. (1996). Corpora are becoming mainstream. In J. Thomas & M. Short (Eds.), *Using corpora for language research* (pp. 3-13). Harlow: Longman.
- Tabata, T. (2002). Investigating stylistic variation in Dickens through Correspondence Analysis of word-class distribution. In T. Saito, J. Nakamura & S. Yamazaki (Eds.), *English corpus linguistics in Japan* (pp. 165-182). Amsterdam: Rodopi.
- Tabata, T. (2004). -ly 副詞の生起頻度解析による文体識別ーコレスポネンス分析と主成分分析による比較研究ー [Discriminating writing styles through frequency of -ly adverbs: Comparison study of correspondence analysis and principal component analysis]. 『電子化言語資料分析研究』 [Research on digitized linguistic resources] (pp.97-114). Osaka: Osaka University Graduate School of Language and Culture Studies.
- Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics*, 4(2), 142-164.

- Taylor, L. (Ed.). (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.
- Taylor, L & Weir, C. J. (Eds.). (2010). *Language testing matters: Investigating the wider social and educational impact of assessment*. New York, NY: Cambridge University Press.
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97, 77-101. doi:10.1111/j.1540-4781.2012.01422.x
- Tono, Y. (2000). A corpus-based analysis of interlanguage development: Analyzing part-of-speech tag sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC '99: Practical applications in language corpora* (pp. 323-340). Frankfurt: Peter Lang.
- Tono, Y. (2006). L2 acquisition of grammatical morphemes. In T. McEnery, R. Xiao, & Y. Tono, *Corpus-based language studies: An advanced resource book* (pp. 247-263). Oxford: Routledge.
- Tono, Y. (2007). (Ed.). 日本人中高生一万人の英語コーパス ”JEFLL Corpus” [A corpus of 10,000 Japanese junior and senior high school students: JEFLL Corpus]. Tokyo, Japan: Shogakukan.
- Tono, Y. (2013). Criterial feature extraction using parallel learner corpora and machine learning. In Díaz-Negrillo, A., Baillier, N., & Thompson, P. (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 169-203). Amsterdam: Benjamins.
- Tono, Y., Kaneta, T., & Doi, Y. (2011, October). CEFR Level Marker の開発 : 英語到達度指標研究のコーパス言語学的展開 [Development of CEFR level marker: Expanding corpus linguistics for investigating indices for English achievement]. Paper presented at the Annual Meeting of Japan Association of Corpus Linguistics, Kyoto.
- Tono, Y., Kawaguchi, Y. & Minegishi, M. (Eds.) (2012). *Developmental and cross-linguistic perspectives in learner corpus research*. Amsterdam: Benjamins.
- Van Rooy, B., & Terblanche, L. (2009). A multi-dimensional analysis of a learner corpus. In A. Renouf & A. Kehoe (Eds.), *Corpus linguistics: Refinements and reassessments* (pp. 239-254). Amsterdam: Routledge.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal*, 97, 11-30. doi:10.1111/j.1540-4781.2012.01421.x

Yelland, P. M. (2010). An introduction to correspondence analysis. *The Mathematica Journal*, 12, 1-23. doi:10.3888/tmj.12-4

## **APPENDICES**

## APPENDIX A

### A SUMMARY OF THE SST ASSESSMENT CRITERIA

The following table was created with the information retrieved on July 11, 2011 from the ALC Press Web site: <http://www.alc.co.jp/edusys/sst/e/index.html>.

Level	Function	Text type
1	<ul style="list-style-type: none"> <li>• Unable to respond to questions</li> <li>• Long silences</li> </ul>	<ul style="list-style-type: none"> <li>• Japanized English words and memorized phrases</li> </ul>
2	<ul style="list-style-type: none"> <li>• Take time to start speaking</li> <li>• Respond to questions but cannot add information</li> </ul>	<ul style="list-style-type: none"> <li>• Japanized English words and short phrases</li> <li>• Other than formulaic expression, many sentences are incomplete</li> </ul>
3	<ul style="list-style-type: none"> <li>• Respond appropriately to a familiar topic</li> <li>• Convey basic matter in simply-structured sentences in 50% success rate</li> </ul>	<ul style="list-style-type: none"> <li>• Words and phrases</li> <li>• Incomplete and simply-structured sentences</li> </ul>
4	<ul style="list-style-type: none"> <li>• Respond to various questions</li> <li>• Can describe and explain reasons</li> <li>• Can add information to a familiar topic</li> </ul>	<ul style="list-style-type: none"> <li>• Simple sentences or compound sentences (e.g., and, but)</li> <li>• Monotonous</li> </ul>
5	<ul style="list-style-type: none"> <li>• In most cases, can add information to a familiar topic</li> <li>• Manage to communicate with familiar vocabulary</li> </ul>	<ul style="list-style-type: none"> <li>• Conscious of logic between sentences</li> <li>• Unnatural use of conjunctions (e.g., and, but, when, because)</li> </ul>
6	<ul style="list-style-type: none"> <li>• Easy to talk about a familiar topic</li> <li>• Can compare and explain past events</li> <li>• Lack of information in a complex topic</li> <li>• Flow of speech stops in a complex topic</li> </ul>	<ul style="list-style-type: none"> <li>• Simple, compound, and complex sentences</li> <li>• Occasional use of if-conditional sentences and relative pronouns</li> </ul>
7	<ul style="list-style-type: none"> <li>• Can add one's own experience, give detailed description, and brief comments to expand the talk</li> </ul>	<ul style="list-style-type: none"> <li>• Simple, compound, and complex sentences</li> <li>• Occasional use of paragraphs</li> </ul>
8	<ul style="list-style-type: none"> <li>• Mostly specific and persuasive</li> <li>• Coming to talk logically</li> </ul>	<ul style="list-style-type: none"> <li>• Basically coherent paragraphs</li> <li>• Some complicated sentence structures cause periphrastic expression</li> </ul>
9	<ul style="list-style-type: none"> <li>• Always logical and persuasive</li> <li>• Structure of talk is easily understandable in an abstract topic</li> </ul>	<ul style="list-style-type: none"> <li>• Appropriate and effective use of various text types</li> <li>• Conscious of structure to develop logical story</li> </ul>

Level	Vocabulary	Grammar
1	<ul style="list-style-type: none"> <li>Lack of basic vocabulary</li> </ul>	<ul style="list-style-type: none"> <li>Lack of basic knowledge</li> <li>Frequent major errors</li> </ul>
2	<ul style="list-style-type: none"> <li>Mostly use proper nouns, loanwords, or basic vocabulary</li> <li>Wrong use of part-of-speech</li> </ul>	<ul style="list-style-type: none"> <li>Likely to construct sentences in Japanese word order</li> </ul>
3	<ul style="list-style-type: none"> <li>Limited use of highly-frequent basic vocabulary</li> <li>Misunderstanding of meaning</li> </ul>	<ul style="list-style-type: none"> <li>Mostly correct simple sentences</li> <li>Wrong word order in long sentences</li> </ul>
4	<ul style="list-style-type: none"> <li>Combine basic vocabulary to construct sentences</li> <li>Inappropriate choice of vocabulary</li> </ul>	<ul style="list-style-type: none"> <li>Unsteady correctly used verbs and tenses confuses listeners</li> <li>Omission of conjunctions (e.g., and, but)</li> </ul>
5	<ul style="list-style-type: none"> <li>Lack of adjectives, adverbs, and idioms</li> <li>Occasional use of advanced words and idioms</li> </ul>	<ul style="list-style-type: none"> <li>Frequent minor errors</li> <li>Wrong tense and word order in difficult sentence structures (e.g., relative clause)</li> </ul>
6	<ul style="list-style-type: none"> <li>Sometimes choose wrong words but understandable</li> </ul>	<ul style="list-style-type: none"> <li>Minor errors and unsteady use of tenses in comparatively difficult expressions and sentence structures</li> </ul>
7	<ul style="list-style-type: none"> <li>Frequent use of advanced vocabulary</li> <li>Rarely use colloquial and formal expressions inappropriately</li> <li>Rarely use unnatural expressions</li> </ul>	<ul style="list-style-type: none"> <li>Good command in using tenses</li> <li>Make small mistakes</li> </ul>
8	<ul style="list-style-type: none"> <li>Native-like natural expression</li> <li>Occasionally use too formal expression</li> </ul>	<ul style="list-style-type: none"> <li>Elementary errors sometimes occur in complex sentences but not habitual</li> </ul>
9	<ul style="list-style-type: none"> <li>Colloquial to technical expression</li> <li>Appropriate choice of vocabulary in an unfamiliar topic</li> </ul>	<ul style="list-style-type: none"> <li>Occasional minor errors</li> <li>Can produce grammatically correct speech unconsciously</li> </ul>

Level	Pronunciation/fluency
1	<ul style="list-style-type: none"> <li>• Long silence</li> </ul>
2	<ul style="list-style-type: none"> <li>• Strong Japanese accent is barely intelligible</li> <li>• Slow speed</li> </ul>
3	<ul style="list-style-type: none"> <li>• Some repetition and rephrasing but a few long pauses</li> <li>• Strong Japanese accent</li> </ul>
4	<ul style="list-style-type: none"> <li>• Consistent Japanese accent sometimes interferes</li> <li>• Still takes time to speak</li> <li>• Many self-corrections sometimes provide understanding</li> </ul>
5	<ul style="list-style-type: none"> <li>• Smooth in a familiar topic but choppy in a difficult topic</li> <li>• Understandable to most native speakers of English</li> </ul>
6	<ul style="list-style-type: none"> <li>• Many self-corrections, repetitions, fillers in a complicated topic and difficult sentences</li> <li>• Japanese accent does not hinder the understanding</li> </ul>
7	<ul style="list-style-type: none"> <li>• Some repetitions and self-corrections in an unfamiliar topic and difficult sentences, but fluent in other occasions</li> <li>• Much less Japanese accent</li> </ul>
8	<ul style="list-style-type: none"> <li>• Mostly smooth and native-like</li> <li>• Overall intonation is slightly unnatural in an unexpected topic</li> </ul>
9	<ul style="list-style-type: none"> <li>• No problems for native speakers of English to understand</li> </ul>

## APPENDIX B

### PART-OF-SPEECH TAGSETS USED FOR TREETAGGER

Tags	Description	Example
CC	coordinating conjunction	and, but, nor, or, yet <u>Yet</u> it's cheap, cheap <u>yet</u> good
	mathematical operator	plus, minus, less times (in the sense of <i>multiplied by</i> ) over (in the sense of <i>divided by</i> )
	for (in the sense of <i>because</i> )	He asked to be transferred, <u>for</u> he was unhappy.
CD	cardinal number	one, two, three
DT	article	a, an, the
	determiner	another, any, some, each, no, every, <u>either</u> way <u>neither</u> decision that, these, this, those
	When determiners are used pronominally, that is, without a head noun, they are tagged as determiners not as common nouns.	I can't stand <u>this</u> . I'll take <u>both</u> . <u>Either</u> would be fine. *DT or NN
	<i>all</i> and <i>both</i> (when they do not precede a determiner or possessive pronoun)	<u>all</u> roads <u>both</u> times *DT or PDT
EX	existential <i>there</i>	<u>There</u> was a party in progress. <u>There</u> ensued a melee.
FW	foreign word	persona non grata
IN	preposition precedes a noun phrase or prepositional phrase	in, of, like
	subordinating conjunction	until, than (comparison)
	subordinating conjunction <i>when</i> (precedes a clause)	I like it <u>when</u> you make dinner for me.
	subordinating conjunction <i>that</i> (when introduces complements of	the fact <u>that</u> you're here the claim <u>that</u> angels have wings

---

	nouns)	*IN or WDT
	so (in the sense of <i>so that</i> )	
JJ	adjective	tall the only <u>such</u> case
	hyphenated compounds	happy-go-lucky
	ordinal numbers which are compounds of the form n-th, X-est	fourth-largest
	adjectives with a comparative meaning without the ending <i>-er</i>	superior
	adjectives with the ending <i>-er</i> without a strictly comparative meaning	further
	adjectives with a superlative meaning without the ending <i>-est</i>	first, last, unsurpassed
JJR	comparative adjective adjective with the comparative ending <i>-er</i> and a comparative meaning	taller more, less (in case they occur by themselves)
JJS	superlative adjective adjective with the superlative ending <i>-est</i> , and worst	tallest, most, least
LS	list item marker (letters and numerals used to identify items in a list)	1), 2), 3)
MD	modal verb	can, could, (dare), may, might, must, ought, shall, should, will, would
NN	common noun, singular or mass	apple
	indefinite pronoun	naught, none, compounds of any-, every- and some- with <i>-one</i> and <i>-thing</i> no-one, no one (no/DT one/NN)
	impersonal pronoun	the only <u>one</u> of its kind collocation (another <u>one</u> ) One of them is good, but the <u>other</u> is bad.

---

---

	nouns in pronominal position that function as modifiers (not adjectives)	A cotton/NN shirt The nearest book/NN store wool/NN sweater (woolen/JJ sweater)
NNS	common noun, plural	apples
NP	proper noun, singular words referring to languages or nations	Tom an English/NP sentence English/JJ cuisine The English/NPS tend to be uninspired cooks. *NP, NPS or JJ
	weekday noun month noun	Wednesday February
NPS	proper noun, plural	Vikings
PDT	predeterminer determinerlike elements when they precede an article of possessive pronoun	<u>half</u> his time <u>many</u> a moon <u>nary</u> a soul <u>quite</u> a mess <u>rather</u> a nuisance <u>such</u> a good time
	<i>all</i> and <i>both</i> when they precede a determiner or possessive pronoun	<u>all</u> his marbles <u>both</u> the girls
	<i>such</i> precedes a determiner	<u>such</u> a good time
POS	possessive ending on nouns ending in 's or '	John's idea the parents' distress
PP	personal pronoun without regard for case distinctions	I, me, you he, him
	reflective pronoun ending in <i>-self</i> or <i>-selves</i>	myself, ourselves
	nominal possessive pronoun	mine, yours, his, hers, ours, theirs
	impersonal third person pronoun	<u>One</u> doesn't do that kind of thing in public. it
PPS	adjectival possessive pronoun	my, your, his, her, its, one's, our, their
RB	Adverb	here, good, however

---

---

	adverb ends in <i>-ly</i>	usually, naturally
	degree words	quite, too, very
	post head modifiers	good <u>enough</u> very well <u>indeed</u>
	negative markers	not, n't, never
RBR	comparative adverb adverbs with the ending <i>-er</i> without a strictly comparative meaning	better We can always come by <u>later</u> .
RBS	superlative adverb	best
RP	particle a number of mostly monosyllabic words that also double as directional adverbs and prepositions	give <u>up</u>
SYM	mathematical, scientific, and technical symbols or expressions that are not words of English	% & ' " " ( ) * + , . < = > @
TO	preposition, infinitival marker	to him, to go
UH	interjection	<u>My</u> , what a gorgeous day. <u>See</u> , it's like this. oh, please, well, yes
VB	be-verb, base form (imperatives)	be
	be-verb, base form (infinitives)	be
	be-verb, base form (subjunctives)	be
VBD	be -verb, past tense, conditional form of the verb <i>to be</i>	were, was If I <u>were</u> rich, ... If I <u>were</u> to win the lottery, ...
VBG	be -verb, gerund or present participle	being
VBN	be -verb, past participle	been
VBP	be -verb, non-3rd person singular present	am, are
VBZ	be -verb, 3rd person singular present	is

---

---

VH	have, base form (imperatives)	have
	have, base form (infinitives)	have
	have, base form (subjunctives)	have
VHD	have, past tense	had
VHG	have, gerund or present participle	having
VHN	have, past participle	had
VHP	have, non-3rd person singular present	have
VHZ	have, 3rd person singular present	has
VV	other verbs, base form (imperatives)	<u>Do</u> it.
	other verbs, base form (infinitives)	You should <u>do</u> it. We want them to <u>do</u> it. We made them <u>do</u> it.
	other verbs, base form (subjunctives)	We suggested that he <u>do</u> it.
VVD	other verbs, past tense	did
VVG	other verbs, gerund, or present participle	doing <u>According</u> to reliable sources <u>Concerning</u> your request of last week
VVN	other verbs, past participle	done <u>Granted</u> that he is coming <u>Provided</u> that he comes
VVP	other verbs, non-3rd person singular present	do
VVZ	other verbs, 3rd person singular present	does
WDT	wh-determiner wh-word precedes a head noun	<u>What</u> kind do you want? I don't know <u>what</u> kind you want. <u>Which</u> book do you like better? I don't know <u>which</u> book you like better. Which do you like better? I don't know which you like better. I'll get you whichever you want.
	which (relative pronoun)	a man <u>that</u> I know

---

---

	that (relative pronoun)	
WP	wh-pronoun (what, who, whom) whatever	Tell me <u>what</u> you would like to eat I'll get you <u>whatever</u> you want.
WP\$	possessive wh-pronoun	Whose
WRB	wh-adverb (e.g., how, where, why)  <i>when</i> in temporal sense	  When he finally arrived, I was on my way out.

---

*Note.* Most of the descriptions and examples are from Santorini (1991).

## APPENDIX C

### TAGS OF TARGETED LINGUISTIC FEATURES

The following notational conventions describe the linguistic features largely extracted from Biber (1988, pp. 222-223).

+: used to separate constituents

( ): marks optional constituents

/: marks disjunctive options

xxx/yyy/zzz: stands for any word

#: marks a word boundary

T#: marks a 'tone unit' boundary, defined in Quirk, Randolph, Greenbaum, Leech, and Svartvik (1985) for use in the London-Lund corpus.

DO: *do, does, did, don't, doesn't, didn't, doing, done*

HAVE: *have, has, had, having, -'ve, -'d, haven't, hasn't, hadn't*

BE: *am, is, are, was, were, being, been, 'm, 're, isn't, aren't, wasn't, weren't*

MODAL: *can, may, shall, will, 'll, could, might, should, would, must, can't, won't, couldn't, mightn't, shouldn't, wouldn't, mustn't*

AUX: MODAL/DO/HAVE/BE/-'s

SUBJPRO: *I, we, he, she, they* (plus contracted forms)

OBJPRO: *me, us, him, them* (plus contracted forms)

POSSPRO: *my, our, your, his, their, its* (plus contracted forms)

REFLEXPRO: *myself, ourselves, himself, themselves, herself, yourself, yourselves, itself*

PRO: SUBJPRO/OBJPRO/POSSPRO/REFLEXPRO/*you/her/it*

PREP: prepositions (e.g., *at, among*) (see no. 33)

CONJ: conjuncts (e.g., *furthermore, therefore*) (see no. 39)

ADV: adverbs

ADJ: adjectives

N: nouns

VCN: any past tense or irregular past participial verb

VBG: *-ing* form of verb

VB: base form of verb

VBZ: third person, present tense form of verb

V: verbs

WHP: WH pronouns – *who, whom, whose, which*

WHO: other WH words – *what, where, when, how, whether, why, whoever, whomever, whichever, wherever, whenever, whatever, however*

ART: articles – *a, an, the*

DEM: demonstratives – *this, that, these, those*

QUAN: quantifiers – *each, all, every, many, much, few, several, some, any*

NUM: numerals – *one ... twenty, hundred, thousand*

DET: ART/DEM/QUAN/NUM

ORD: ordinal numerals – *first ... tenth*

QUANPRO: quantifier pronouns – *everybody, somebody, anybody, everyone, someone, anyone, everything, something, anything*

TITLE: address titles – *mr, mrs, mis, ms, dr*

CL-P: clause punctuation (‘.’, ‘!’, ‘?’, ‘:’, ‘;’, ‘-’)

ALL-P: all punctuation (CL-P plus ‘,’)

The following is a detailed explanation of the linguistic features used in present study. They largely replicate the work of Biber (1988) and Murakami's (2009) revised version of Biber (1988).

1. past tense

*Note 1:* Any word that is part-of-speech tagged as a past tense (VBD, VHD and VVD).

2. perfect aspect

(a) HAVE + (ADV) + (ADV) + VBN/VBD

(b) HAVE + N/PRO + VBN/VBD

*Note 1:* As past participles were recognized as past tenses (63 cases out of 1807 cases), past tense "VBD" tag was added to Biber (1988) as in Murakami (2009).

3. present tense

All VB (base form) or VBZ (third person singular present) verb forms in the dictionary, excluding infinitives.

4. place adverbials

aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, nowhere, outdoors, outside, overboard, overland, overseas, underfoot, underground, underneath, uphill, upstairs, upstream, east, north, south, west

5. time adverbials

afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, soon, subsequently, today, tomorrow, tonight, yesterday

6. first person pronouns

I, my, me, mine, myself, we, our, us, ours, ourselves

*Note 1:* Possessive pronouns, *mine* and *ours* were added to the Biber (1988) as in Murakami (2009).

7. second person pronouns

you, your, yours, yourself, yourselves

*Note 1:* Possessive pronouns, *yours* was added to Biber (1988) as in Murakami (2009).

8. third person pronouns

he, his, him, himself, she, her, hers, herself, they, their, them, theirs, themselves

*Note 1:* Possessive pronouns, *hers* and *theirs* were added to Biber (1988) as in Murakami (2009).

9. pronoun *it*

it, its, itself

*Note 1:* *Its* and *itself* were added to Biber (1988).

#### 10. demonstrative pronouns

that, this, these, those

that/this/these/those + V/AUX/CL-P/WHP/and (where *that* is not a relative pronoun)

*Note 1:* In Biber (1988) T# was added as a restriction, but it was excluded here.

#### 11. indefinite pronouns

anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, nowhere, somebody, someone, something

#### 12. pro-verb *do*

DO in all cases except for the following:

(a) DO + (ADV) + V (DO as auxiliary)

(b) ALL-P/T#/WHP + DO (where DO is not *-ing* form) (DO as question)

*Note 1:* The condition that “DO is not *-ing* form” was added to Biber (1988) in algorithm (b), because sentence initial *doing* can function as a pro-verb as suggested in Murakami (2009).

#### 13. direct WH-questions

ALL-P/T#+WHO/WHP+AUX/V

*Note 1:* WHP and V were added to Biber (1988), because it helps to retrieve *wh* questions better as suggested in Murakami (2009). According to Murakami (2009), the original algorithm in Biber (1988) failed to retrieve some question sentences (e.g., *Who went there?*).

*Note 2:* All punctuation was used instead of clause punctuation in Biber (1988).

#### 14. nominalizations

All words ending in *-tion*, *-ment*, *-ness*, or *-ity* (plus plural forms)

*Note 1:* There were two problems in counting total frequency of nominalization. First, the programming script was designed to search words ending with special suffixes that represent nominalization, but some words that end with suffixes that stand for nominalization cannot be categorized as nominalization (e.g., *business*). Therefore, the following words that appear in the corpus were deleted from total frequency counts: *city, cities, commodity, pity, quality, university, opportunity, opportunities, emotion, condition, fiction, lotion, mention, nation, nonfiction, portion, question, station, vacation, auction, audition, friction, function, malfunction, occupation, tradition, tuition, business, tipness, moment, comment, document, apartment, element, garment, instrument, pavement, pigment, monument*. Second, it was difficult to determine whether some words were nominalized or not. These controversial words were checked manually based on the definitions in Carter and McCarthy (2006): nouns formed from verbs or adjectives (e.g., *fly - flight, bright - brightness*) can be regarded as examples of nominalization.

#### 15. total other nouns

All nouns (NN/ NS) and proper nouns (NP/ NPS) excluding pronouns (PP/ PPS).

*Note 1:* Frequencies of nominalizations were excluded manually from total other nouns counts.

#### 16. agentless passives

Subtract *by*-passives from the following:

(a) BE + (ADV) + (ADV) + VBN/VBD

(b) BE + N/PRO + (ADV) + VBN/VBD (question sentence)

*Note 1:* As Biber (1988) failed to capture some sentences (e.g. *Why are you so surprised?*), an optional adverb was added to (b) as suggested in Murakami (2009).

*Note 2:* As past participles were frequently recognized as past tense, past tense (VBD) was added to Biber (1988) as in Murakami (2009).

17. *by*-passives

(a) BE + (ADV) + (ADV) + VBN/VBD + (xxx) + *by*

(b) BE + N/PRO + (ADV) + VBN/VBD + (xxx) + *by* (question sentence)

*Note 1:* The same modifications as in 15 were made here. Additionally, an any-word option (xxx) was added mainly to retrieve phrasal verbs as suggested in Murakami (2009).

18. *be* as main verb

BE + any possessive form/PREP/ADJ

*Note 1:* As suggested in Murakami (2009), DET and TITLE were removed from Biber (1988). If they are left in the algorithm, it is difficult to contrast a clause with a phrase in some cases (e.g., *the big house* versus *the house is big*). According to examples in Murakami (2009), we can rephrase “*the book is his*” as “*his book*,” “*a cup is on the table*” into “*a cup on the table*,” and “*the task is easy*” into “*the easy task*,” but we cannot rephrase “*a dolphin is an animal*” or “*he’s Dr. Lee*” in the same manner.

19. existential *there*

(a) *there* + (xxx) + BE (BE includes future tense and perfect aspect in this case)

(b) *there’s/there’re*

*Note 1:* Future tense and perfect aspect were added to (a) as in Murakami (2009).

*Note 2:* *There’re* was added to (b) as in Murakami (2009).

20. *that* verb complements

(a) *and/nor/but/or/also/ALL-P+ that* +DET/PRO/*there*/plural noun/proper noun/TITLE

(b) V (but not BE) + *that* + xxx + yyy + zzz (where xxx is not V/AUX/CL-P/*and*, and either yyy or zzz is AUX/V)

*Note 1:* In order to avoid the demonstrative use of *that*, yyy and zzz were added to algorithm (b) as in Murakami (2009).

*Note 2:* The algorithm (c) in Biber (1988) were not used in this analysis. The algorithm (c) V (but not BE) + PREP + xxx + N + *that* (where xxx is any number of words, but not = N) was supposed to retrieve a *that* clause preceded by a prepositional phrase (e.g. *you’ll have to explain in words that*), but according to Murakami (2009), this formula captures an unnecessary *that* clause that functions as a relative clause (e.g. *my eyes fell on a locker that read “Stop”*). As the number of matches in algorithm (c) was much smaller than (a) and (b), they were excluded.

*Note 3:* The algorithm (d) T# + *that* was prepared for spoken text in Biber (1988). However, it was not used in this analysis, because it is necessary to check all contexts manually to distinguish *that* complement, relative pronouns, demonstrative pronouns and subordinators.

21. *that* adjective complements

ADJ + *that*

*Note 1:* In Biber (1988) T# was added as a restriction. However, it was excluded in this study because it is necessary to edit complements across intonation boundaries by hand.

22. WH-clauses

V (but not BE) + WHP/WHO + xxx (where xxx is not ALL-P/*to*)

*Note 1:* In Biber (1988) xxx represented any word but AUX to avoid extracting WH questions, but according to Murakami (2009), this restriction also avoid retrieving sentences such as, *I knew who could go there*. Therefore, xxx was defined as any word excluding ALL-P and *to* as in Murakami (2009).

23. *to*-clause

*to* + (ADV) + VB

24. past participial postnominal (reduced relative) clauses

N/QUANPRO + VBN + PREP/V/ADV

*Note 1:* VBN was mostly replaced by VBN/VBD in this study, but it was not replaced in this case. As suggested in Murakami (2009), if VBN is replaced with VBN/VBD here, it will also match unnecessary simple-past-tense sentences.

*Note 2:* In Biber (1988) BE was used instead of V, but as noun phrases including past participial postmodification can be followed by any verb, BE should be replaced by V as suggested in Murakami (2009).

25. *that* relatives on subject position

N + *that* + (ADV) + AUX/V

*Note 1:* In Biber (1988) T# was added as a restriction in this algorithm. However, it was excluded in this study because it is necessary to check *that* relatives across intonation boundaries by hand.

26. *that* relatives on object position

N (not personal pronouns) + *that* + DET/SUBJPRO/POSSPRO/*it*/ADJ/N/TITLE

*Note 1:* Personal pronouns were removed from Biber (1988) as in Murakami (2009). They usually do not function as antecedents in relative clause sentences, and they will only extract unintended sentences (e.g., *I told him that ...*).

*Note 2:* Biber (1988) has restricted nouns to plural nouns, proper nouns, and possessive nouns, but as these restrictions extract unintended matches (e.g., mass nouns), they were not included here as in Murakami (2009).

*Note 3:* In Biber (1988) T# was added as a restriction. However, it was excluded in this study, because according to Biber (1988) *that* relatives sometimes span two intonation units, and it is necessary to exclude this type of unnecessary match manually.

27. WH relatives in subject position

xxx + yyy + N + WHP (but not *whose*) + (ADV) + AUX/V (where xxx and yyy are not any form of the verbs *ask/tell*; to exclude indirect WH questions such as *Tom asked the man who went to the store*)

*Note 1:* A restriction that WHP should be not *Whose* was added as in Murakami (2009). *Whose* does not function as a WH relative in subject position.

*Note 2:* As suggested in Murakami (2009), it is appropriate to add not only xxx as in Biber (1988) but also yyy to exclude sentences as *Tom asked him who went to the store*.

28. WH relatives in object position

xxx + yyy + N + WHP (but not *whose*) + zzz (where xxx and yyy are not any form of the verbs *ask/tell*; to exclude indirect WH questions, and zzz is not ADV/AUX/V, to exclude relativization on subject position)

*Note 1:* Same as number 26.

*Note 2:* Same as number 26.

29. WH relatives with fronted preposition

PREP + WHP (where PREP is lowercase)

*Note 1:* In order to avoid matching an interrogative use of *whom* (e.g. *To whom ...?*), PREP was restricted to lowercase as in Murakami (2009).

30. causative adverbial subordinators

*because*

*Note 1:* As other causative adverbial subordinators (e.g., *as, for, since*) have a range of functions, the adverbial subordinator *because*, which unambiguously functions as a causative adverbial, was only targeted here.

31. concessive adverbial subordinators

although, though

32. conditional adverbial subordinators

if, unless

33. other adverbial subordinators

since, while, whilst, whereupon, whereas, whereby, such that, so that xxx, such that xxx, inasmuch as, forasmuch as, insofar as, insomuch as, as long as, as soon as (where xxx is not N/ADJ)

34. prepositional phrases

against, amid, amidst, among, amongst, at, besides, between, by, despite, during, except, for, from, in, into, minus, notwithstanding, of, off, on, onto, opposite, out, per, plus, pro, re, than, through, throughout, thru, to, toward, towards, upon, versus, via, with, within, without

35. attributive adjectives

Any adjectives not identified as predicative adjectives in number 35.

36. predicative adjectives

(a) BE + ADJ + (xxx) (where xxx is not ADJ/ADV/N)

(b) BE + ADJ + ADV + (xxx) (where xxx is not ADJ/N)

*Note 1:* In Biber (1988) xxx could not be absent, but there are some cases in which punctuation follows (e.g. *it's easy*.) xxx. Therefore, xxx should be allowed to be absent as in Murakami (2009).

37. total adverbs

Conjuncts, hedges, emphatics, discourse particles, downtoners, amplifiers, place adverbials, and time adverbials were excluded from the count of total adverbs.

*Note 1:* The counts of conjuncts and discourse particles were not excluded in Biber (1988), but they should be excluded. They tend to be tagged as adverbs as pointed out in Murakami (2009).

38. conjuncts

alternatively, consequently, conversely, eg, e.g., furthermore, hence, however, i.e., instead, likewise, moreover, namely, nevertheless, nonetheless, notwithstanding, otherwise, rather, similarly, therefore, thus, viz., in comparison, in contrast, in particular, in addition, in conclusion, in consequence, in sum, in summary, in any event, in any case, in other words, for example, for instance, by contrast, by comparison, as a result, as a consequence, on the contrary, on the other hand, that is/else/altogether + , rather + ,/xxx (where xxx is not ADJ/ADV)

*Note 1:* In Biber (1988) T# was added as a restriction. However, it was excluded in this study, because conjuncts are more common in written texts and T# is especially effective in spoken texts.

39. downtoners

barely, hardly, merely, mildly, nearly, partially, partly, practically, scarcely, slightly, somewhat

*Note 1:* *Only* was excluded from the list of Biber (1988). It has many different functions and Japanese EFL learners did not always use *only* as a downtoner such as in cases (e.g. half hour *only* English, I like driving *only*).

40. hedges

at about, something like, more or less, almost, maybe, xxx sort of, xxx kind of (where xxx is not DET/ADJ/POSSPRO/WHO/which)

*Note 1:* *Which* was added to Biber (1988) to avoid matching unintended expressions (e.g. *which kind/sort of*) as in Murakami (2009).

#### 41. amplifiers

absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very

*Note 1:* Among the amplifiers, *very* dominated most of the tokens in the category of amplifiers (*very*: 10,131 tokens, amplifiers except *very*: 196 tokens). Japanese EFL learners seem to have a tendency to overuse this word as a part of chunks such as in “thank you very much” or “very good.”

#### 42. emphatics

for sure, a lot, such a, such an, real + ADJ, so + ADJ, DO + V, just, really + xxx, most, more

*Note 1:* As in Murakami (2009), xxx was added after *really*. If there is no word after it, *really* is commonly used as an expression of surprise, but not as an emphatic.

*Note 2:* As in Murakami (2009), *such an* was added to Biber (1988).

#### 43. discourse particles

(a) ALL-P + well/now/anyway/anyhow/anyways

(b) ALL-P + Well/Now/Anyway/Anyhow/Anyways

*Note 1:* Biber (1988) restricted punctuation to clause punctuation, but the comma was added in algorithm (a).

*Note 2:* Condition (b) was added to Biber (1988) to match sentence-initial target words as in Murakami (2009).

#### 44. possibility modals

can, may, might, could

#### 45. necessity modals

ought, should, must

#### 46. predictive modals

will, would, shall, 'd

*Note 1:* The contraction 'd is ambiguous, because it can indicate either *had* or *would*. However 'd was considered as *would* when it was neither judged as *had* by TreeTagger nor followed by *better* or by past participles as in Murakami (2009).

#### 47. public verbs

acknowledge, admit, agree, assert, claim, complain, declare, deny, explain, hint, insist, mention, proclaim, promise, protest, remark, reply, report, suggest, swear, write

*Note 1:* *say* dominated the tokens in the category of public verbs (*say*: 2,360 tokens, other public verbs: 4,194 tokens). This domination exceeded 30% of the whole tokens, so that *say* was deleted from the public verbs in the list of Biber (1988).

#### 48. private verbs

anticipate, assume, believe, conclude, decide, demonstrate, determine, discover, doubt, estimate, fear, feel, forget, guess, hear, hope, imagine, imply, indicate, infer, know, learn, mean, notice, prove, realize, recognize, remember, reveal, see, show, suppose, think, understand

*Note 1:* *think* was frequently used under the category of private verbs (*think*: 4,633 tokens, other private verbs: 18,942 tokens), but it was not deleted from the private verbs in the list of Biber (1988). It was used to express intellectual states of test-takers.

49. suasive verbs

agree, arrange, ask, beg, command, decide, demand, grant, insist, instruct, ordain, pledge, pronounce, propose, recommend, request, stipulate, suggest, urge

*Note 1:* *ask* was frequently used in the category of suasive verbs (*ask*: 865 tokens, other suasive verbs: 3,481 tokens). This domination did not exceed 30% of the whole tokens and unlike the use of *think* in the category of private verbs, *ask* was not always used in the context of persuading somebody. However, it was excluded from the list of Biber (1988) in the present study.

50. *seem* and *appear*

seem, appear

51. contractions

(a) *n't*

(b) Excluding apostrophes which are used to indicate possessive forms.

*Note 1:* In Biber (1988) T# was added as a restriction, but it was excluded here.

52. stranded prepositions

PREP + ALL-P

53. split infinitives

to + ADV + (ADV) + VB

54. split auxiliaries

AUX + ADV + (ADV) + V (where the first ADV is *not*)

*Note 1:* If the last verb was restricted to VB as in Biber (1988), it will fail to match some sentences (e.g. *they are objectively shown to...*). Therefore, restriction was removed and VB was replaced with V as in Murakami (2009).

*Note 2:* The first ADV cannot be *not* in this algorithm, because if *not* follows AUX immediately, it will match simple negation sentences. As pointed out in Murakami (2009), *not* is usually tagged as an adverb. Thus, the first ADV should be restricted to adverbs other than *not*.

55. phrasal coordination

xxx1 and xxx2 (where xxx1 and xxx2 are both ADV/V/personal pronouns/N)

*Note 1:* In Biber (1988), personal pronouns and other nouns were combined into N, but this matches unnecessary sentences (e.g. *I called Lorri every night and we agonized together*) as pointed out in Murakami (2009). Thus, N was divided into personal pronouns and other nouns.

*Note 2:* In Biber (1988), adverbs were included in the algorithm, but there were only 178 occurrences for ADV and ADV phrasal coordination out of total 4,917 phrasal coordination occurrences. Additionally, ADV and ADV pattern matches more unnecessary sentences (e.g., *there and quickly, somewhere and maybe*) than necessary sentences (e.g., *back and forth, kindly and gently*). Therefore, ADV and ADV phrasal coordination was excluded in this analysis.

56. independent clause coordination

(a) , + *and* + *it/so/then/you/there*+BE/demonstrative pronoun/SUBJPRO

(b) CL-P + *and*

(c) *and* + WHP/WHO/adverbial subordinator/discourse particle/conjunct

*Note 1:* Regarding (a), #T was excluded from the algorithm in Biber (1988). The total frequency counts did not differ much when it was excluded. Thus, it is best to avoid the risk of accepting unnecessary matches.

*Note 2:* Regarding (c), #T was excluded from the restriction in discourse particle. The total frequency counts did not differ much when it was excluded. Thus, it is best to avoid the risk of accepting unnecessary matches.

57. synthetic negation  
(a) *no* + QUANT/ADJ/N  
(b) *neither, nor*

58. analytic negation  
*not, cannot, n't*

## APPENDIX D

### R SCRIPTS FOR CORRESPONDENCE ANALYSIS

```
# Correspondence Analysis
dat <- read.csv(choose.files(), row.names = 1, header = T)

library(anacor)
anacor.result <- anacor(dat, scaling = c("standard", "standard"))
summary(anacor.result)

plot(anacor.result, plot.type = "jointplot")
```

## APPENDIX E

### BOX-AND-WHISKER PLOTS OF TARGETED LINGUISTIC FEATURES

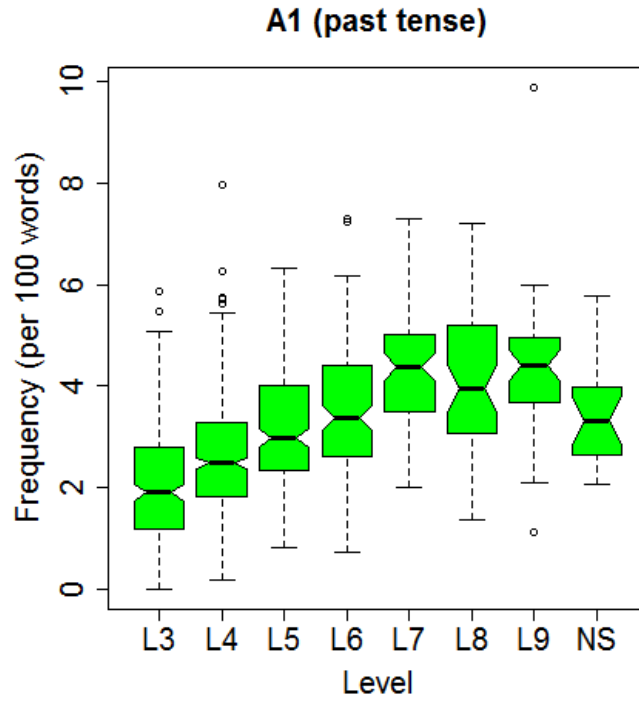


Figure E1. Box-and-whisker plots for A1 (past tense).

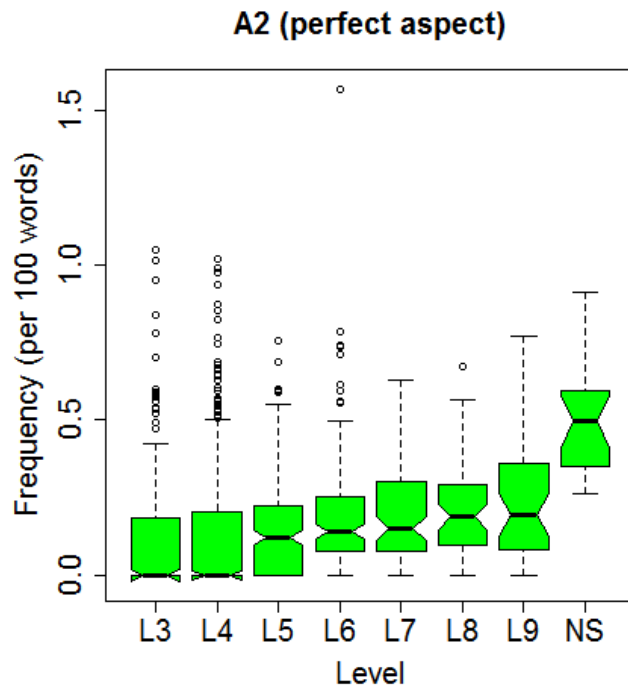


Figure E2. Box-and-whisker plots for A2 (perfect aspect).

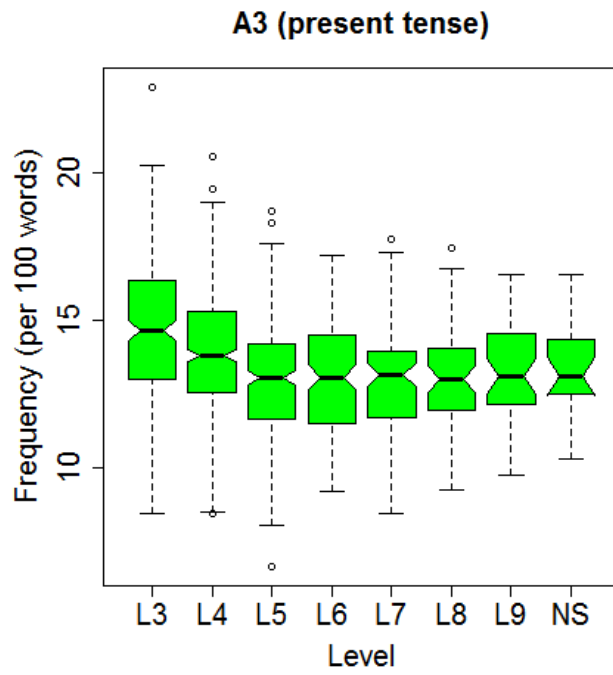


Figure E3. Box-and-whisker plots for A3 (present tense).

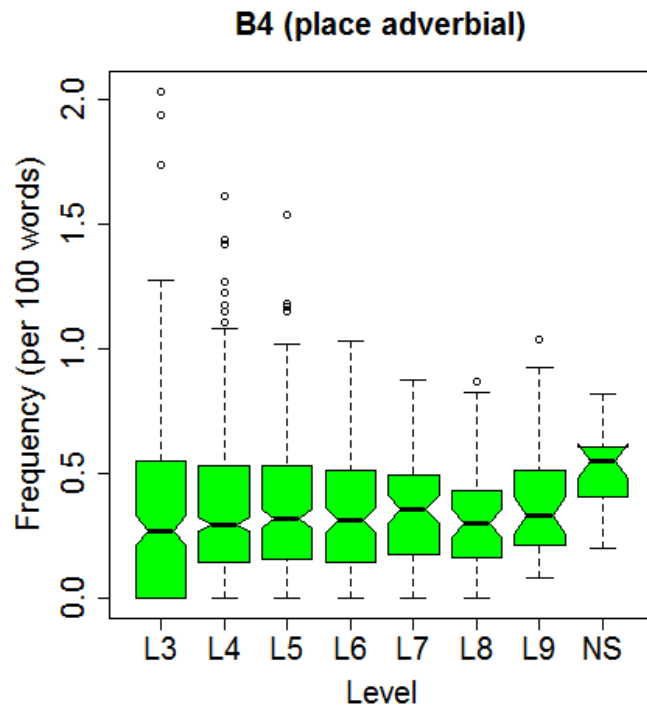


Figure E4. Box-and-whisker plots for B4 (place adverbial).

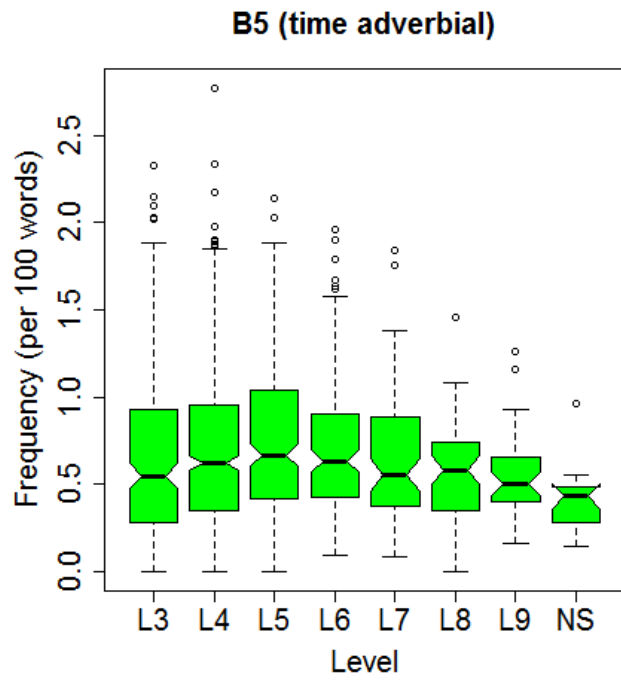


Figure E5. Box-and-whisker plots for B5 (time adverbial).

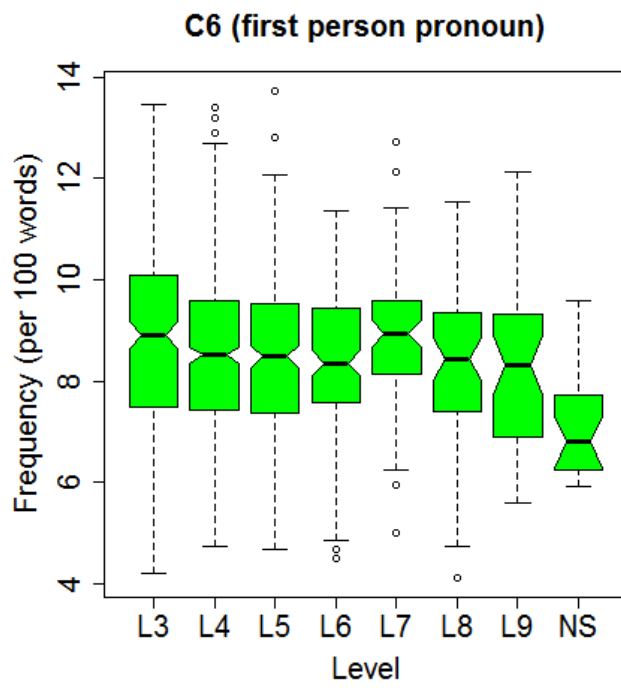


Figure E6. Box-and-whisker plots for C6 (first person pronoun).

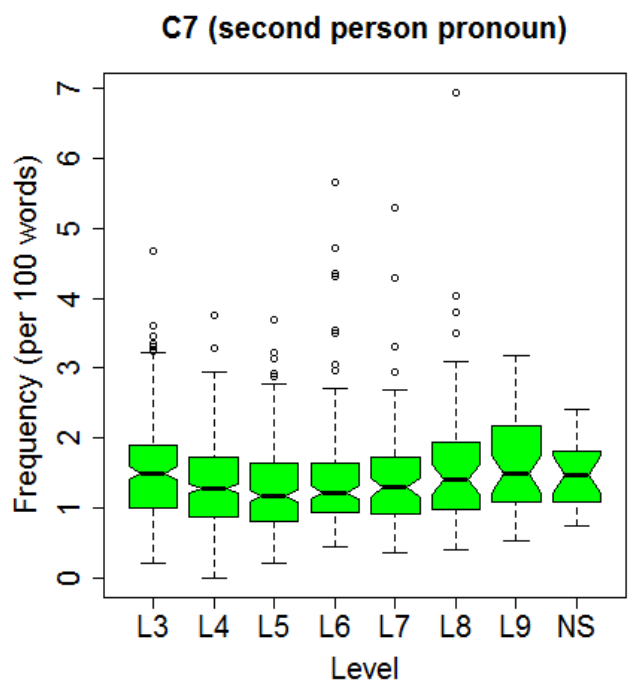


Figure E7. Box-and-whisker plots for C7 (second person pronoun).

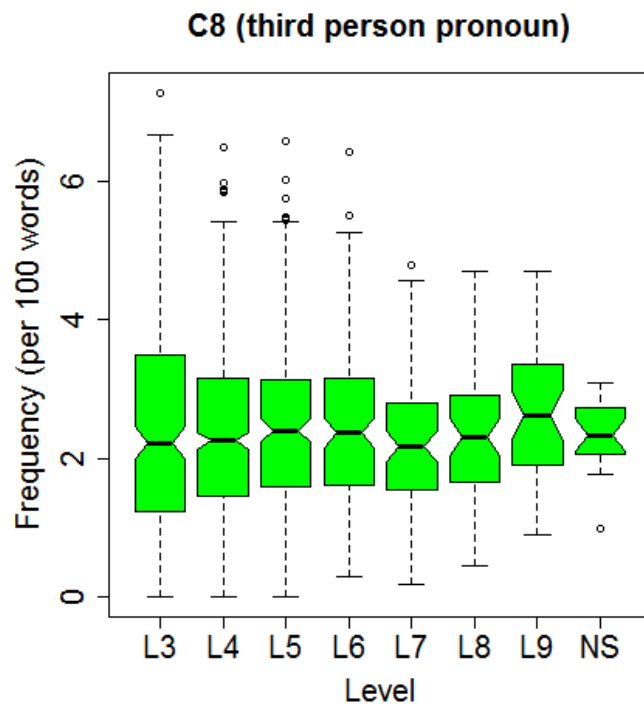


Figure E8. Box-and-whisker plots for C8 (third person pronoun).

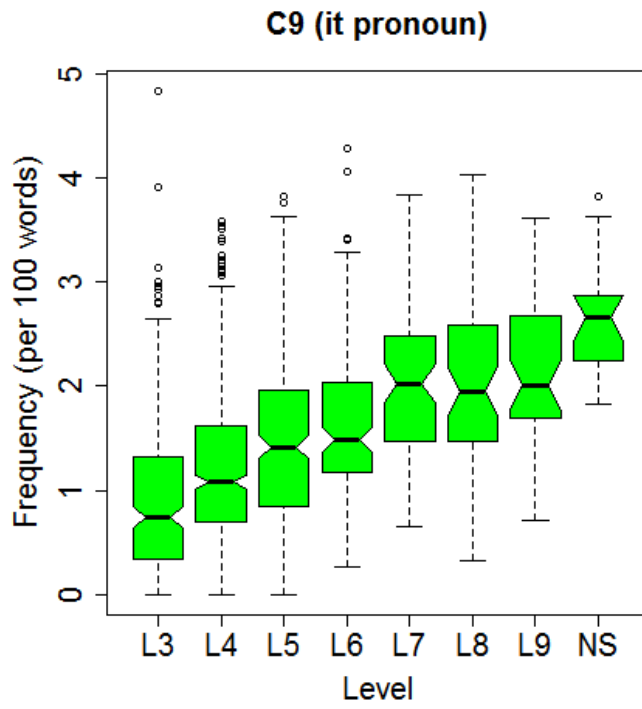


Figure E9. Box-and-whisker plots for C9 (it pronoun).

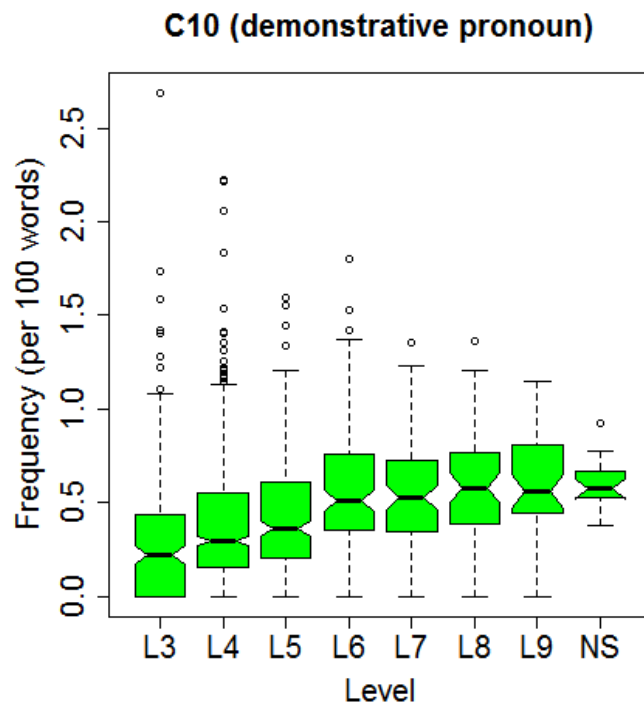


Figure E10. Box-and-whisker plots for C10 (demonstrative pronoun).

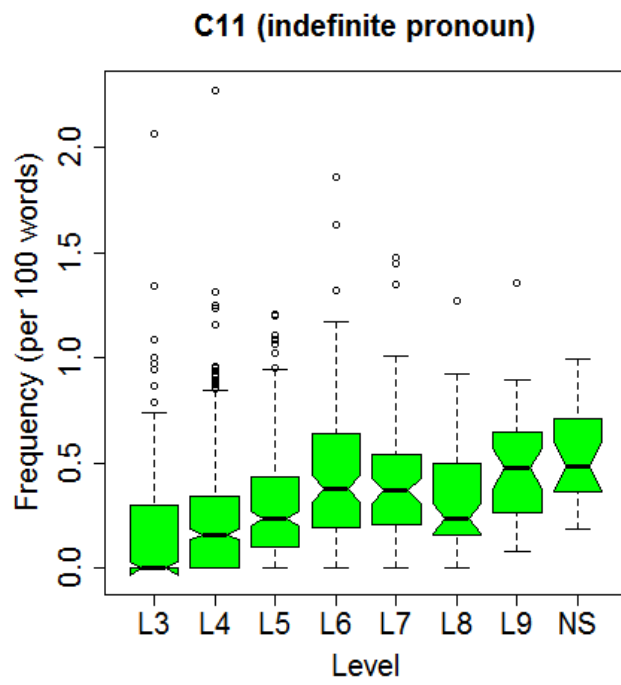


Figure E11. Box-and-whisker plots for C11 (indefinite pronoun).

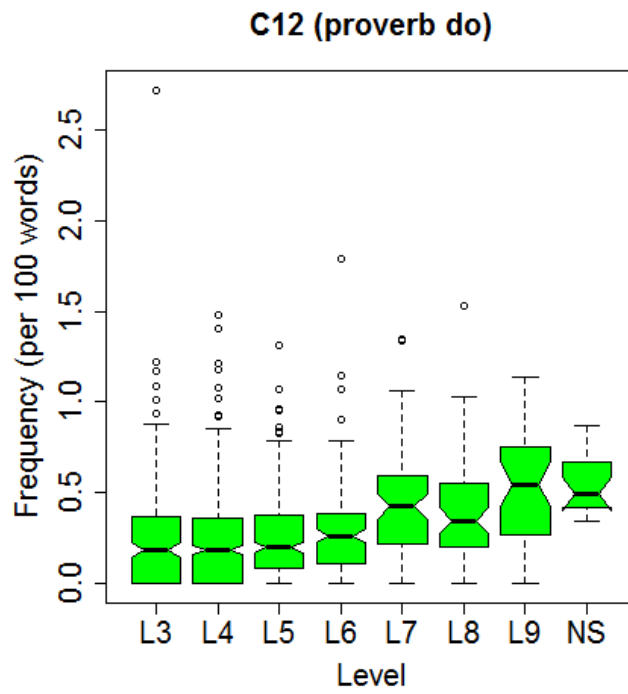


Figure E12. Box-and-whisker plots for C12 (proverb do).

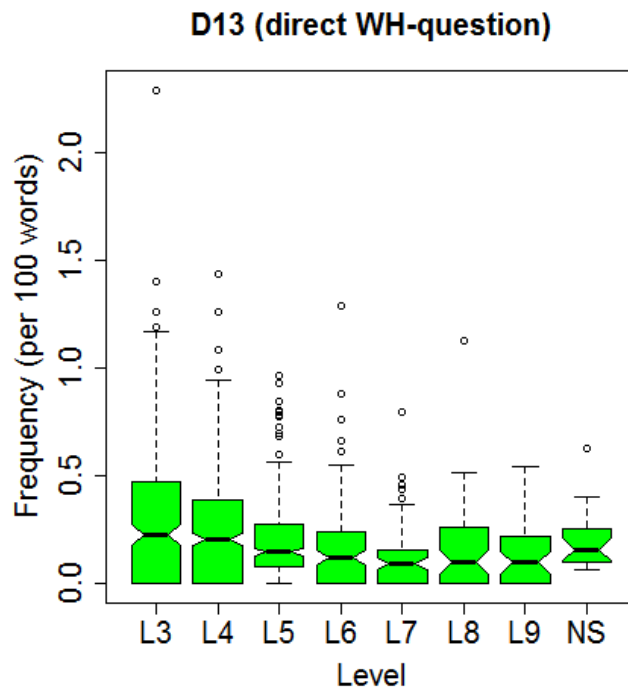


Figure E13. Box-and-whisker plots for D13 (direct WH-question).

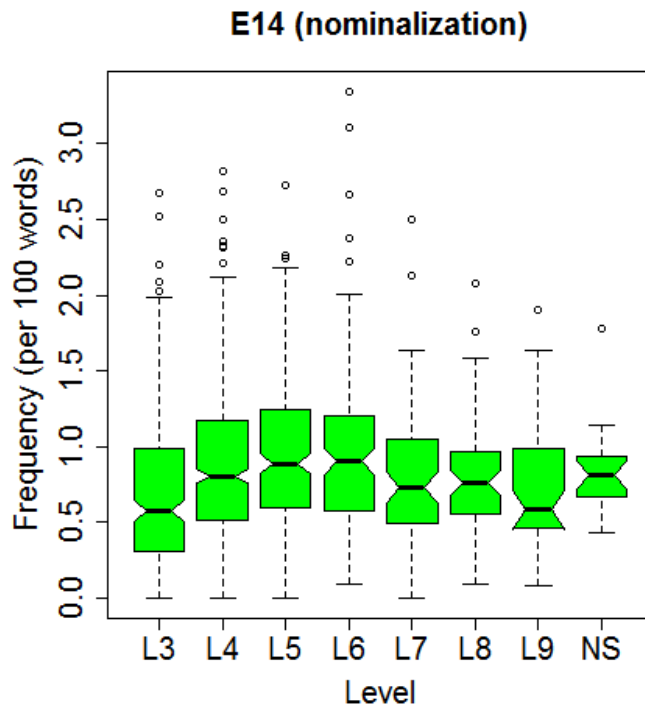


Figure E14. Box-and-whisker plots for E14 (nominalization).

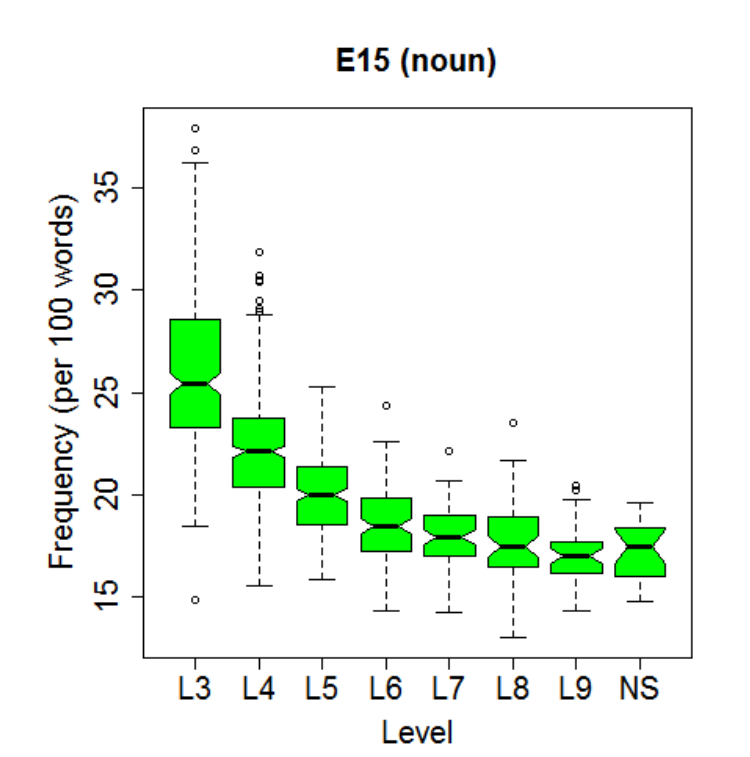


Figure E15. Box-and-whisker plots for E15 (noun).

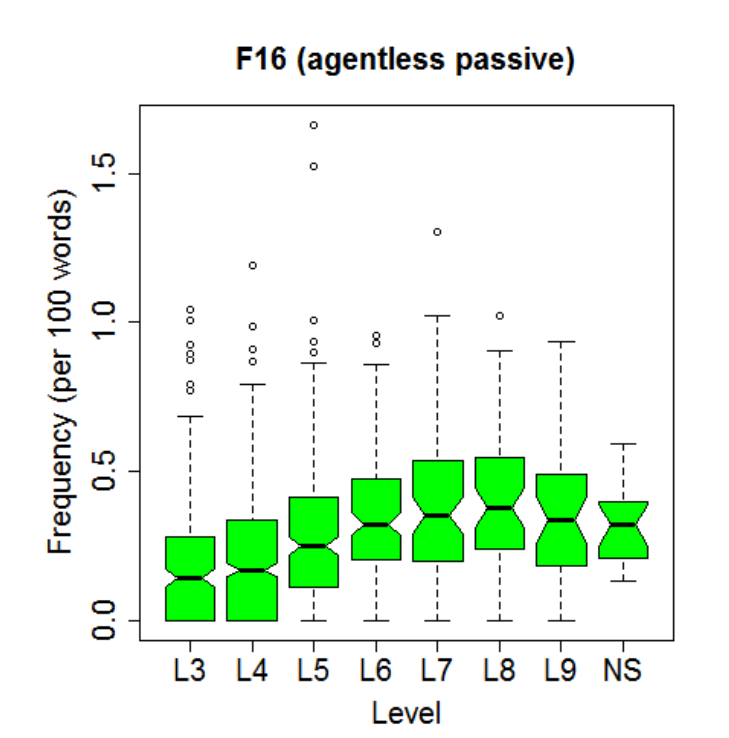


Figure E16. Box-and-whisker plots for F16 (agentless passive).

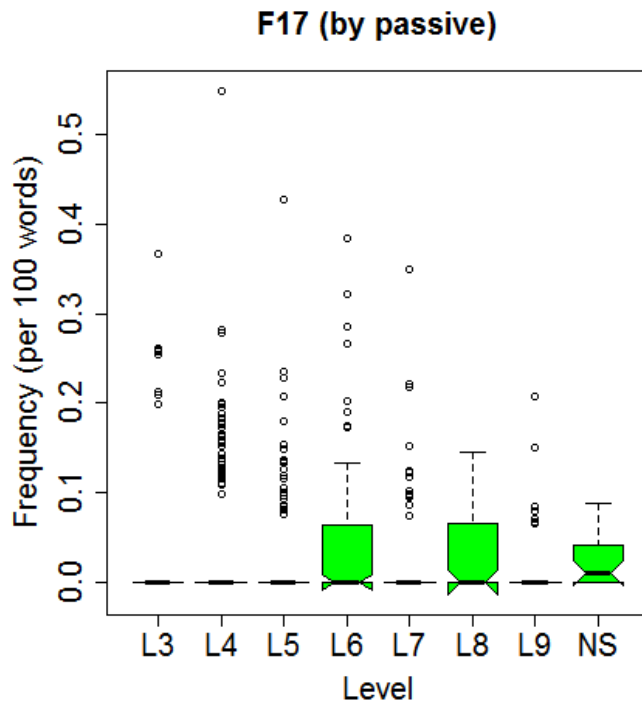


Figure E17. Box-and-whisker plots for F17 (by passive).

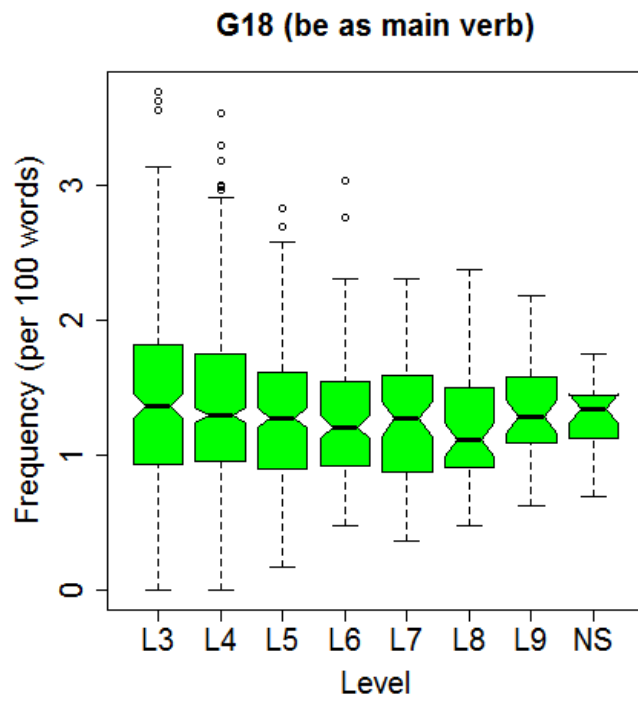


Figure E18. Box-and-whisker plots for G18 (*be* as main verb).

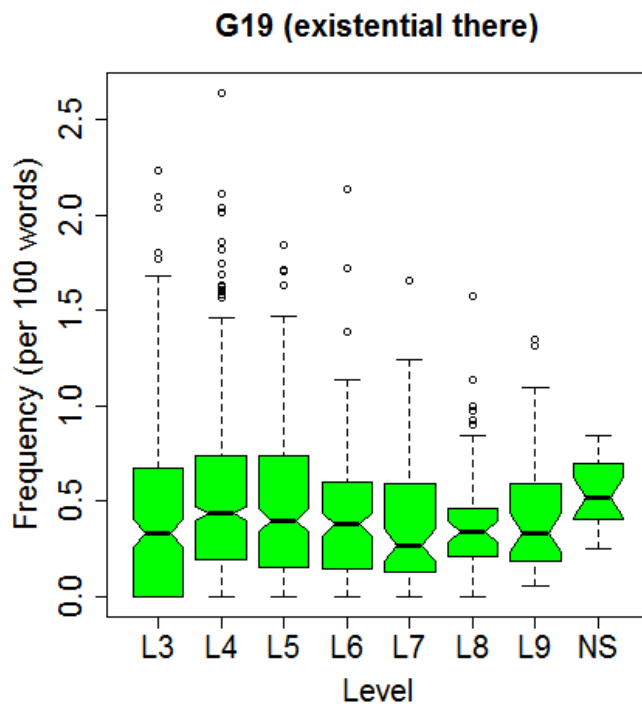


Figure E19. Box-and-whisker plots for G19 (existential there).

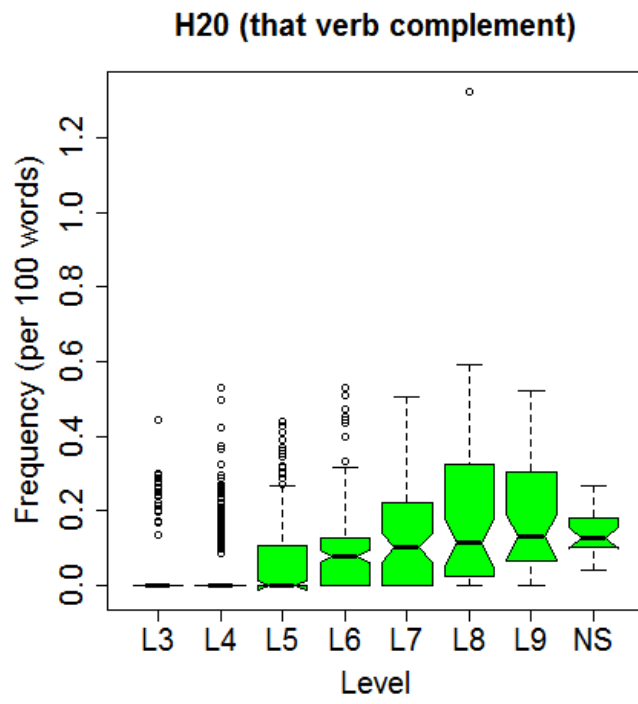


Figure E20. Box-and-whisker plots for H20 (*that* verb complement).

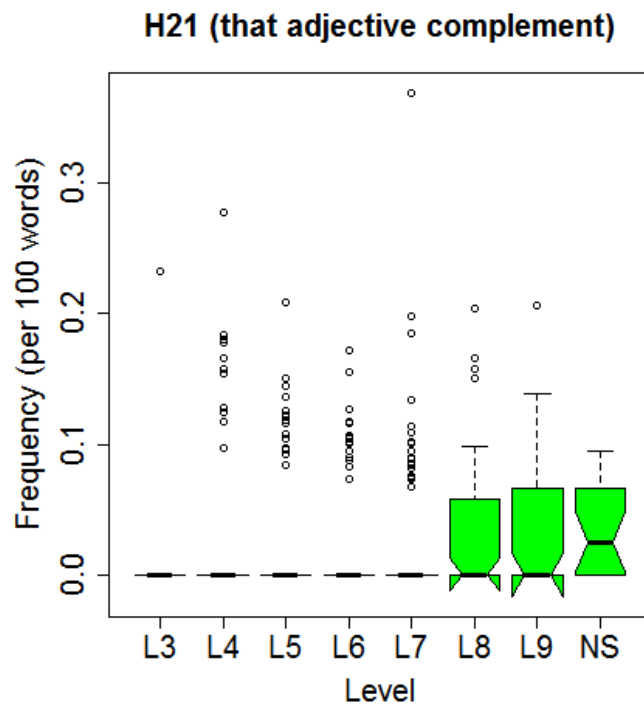


Figure E21. Box-and-whisker plots for H21 (*that* adjective complement).

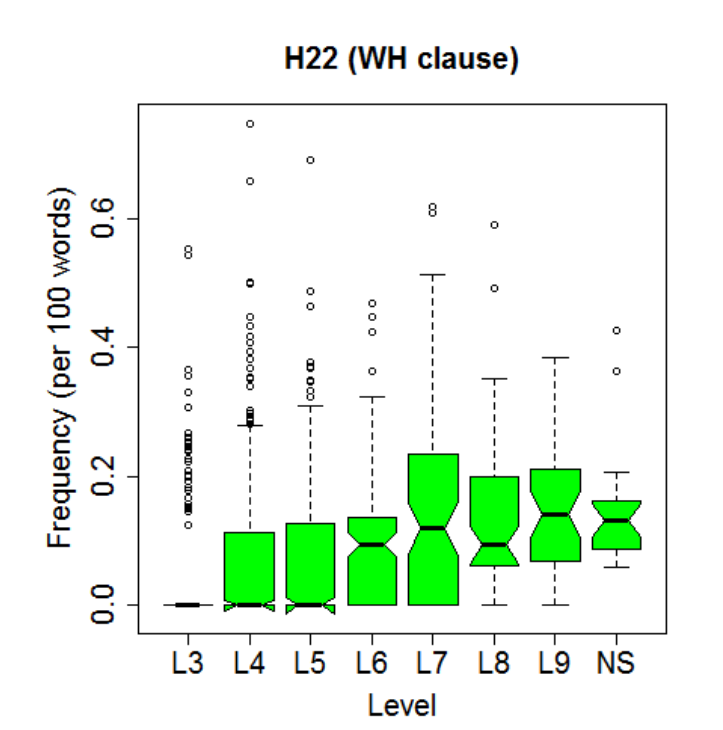


Figure E22. Box-and-whisker plots for H22 (WH clause).

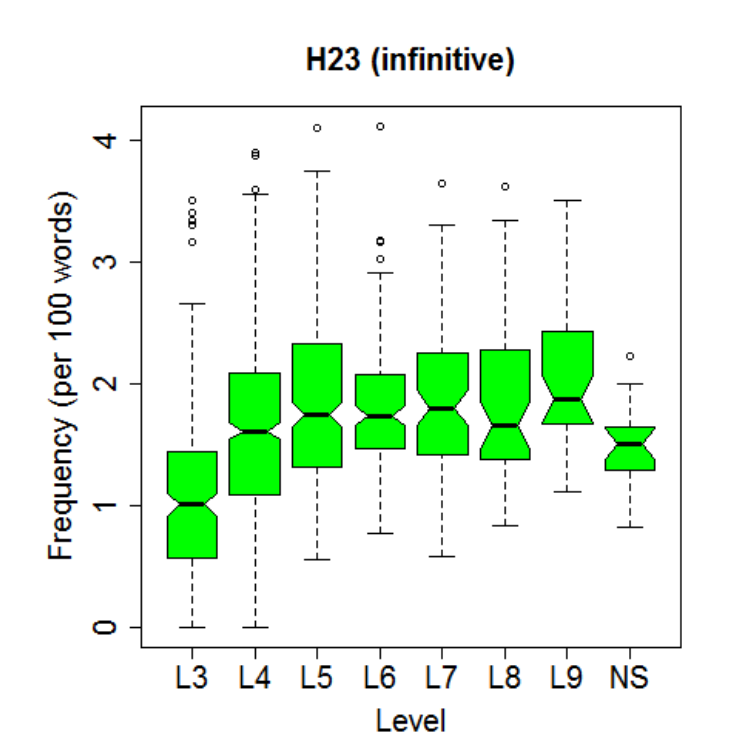


Figure E23. Box-and-whisker plots for H23 (infinitive).

### H24 (past participial postnominal clause)

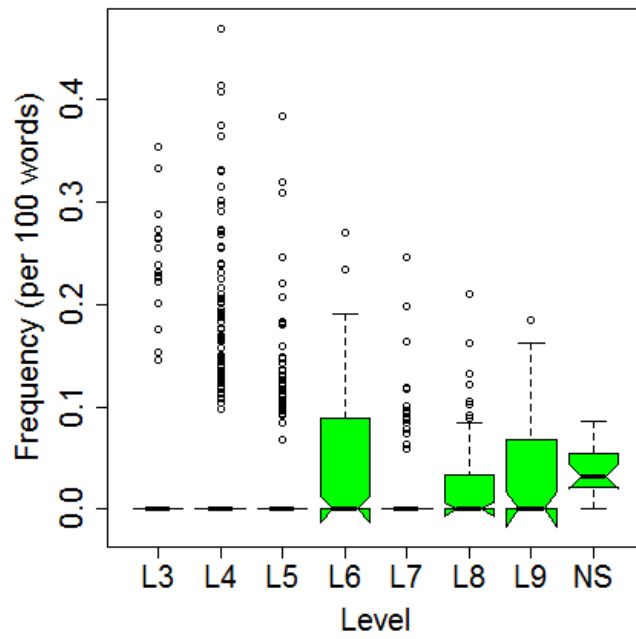


Figure E24. Box-and-whisker plots for H24 (past participial postnominal clause).

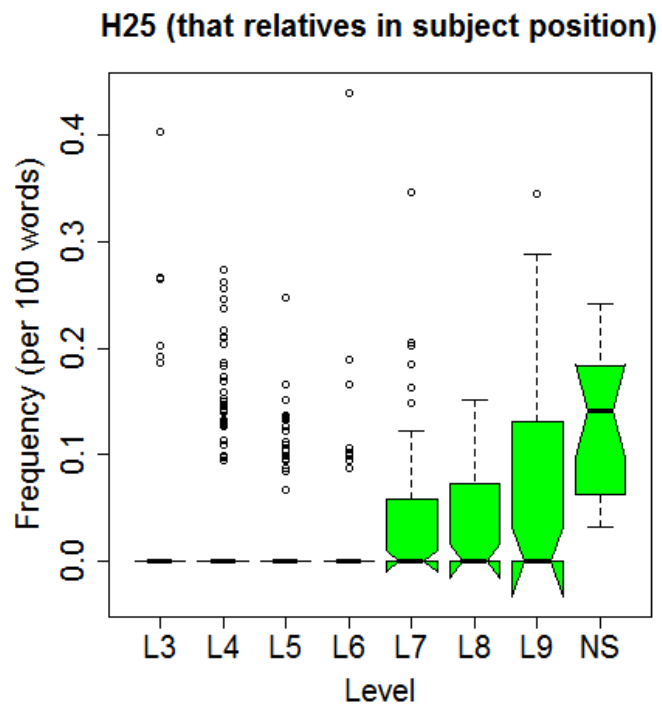


Figure E25. Box-and-whisker plots for H25 (*that* relatives in subject position).

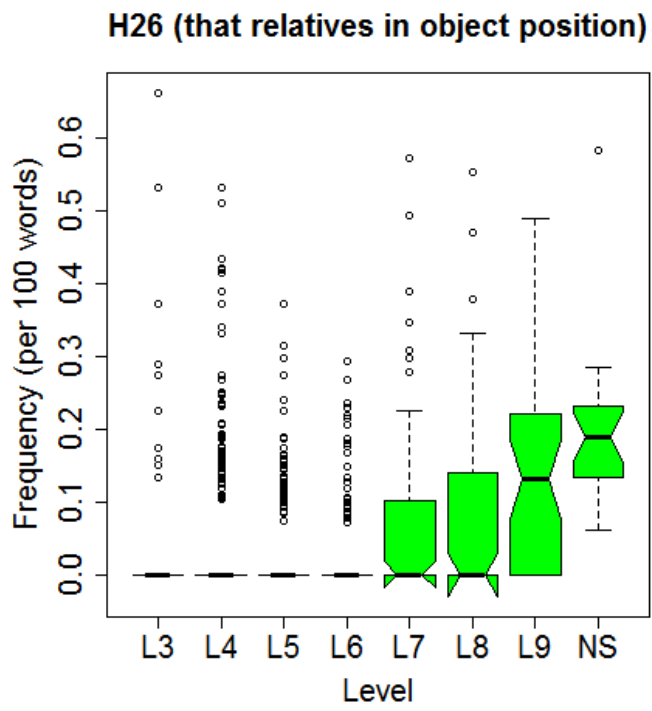


Figure E26. Box-and-whisker plots for H26 (*that* relatives in object position).

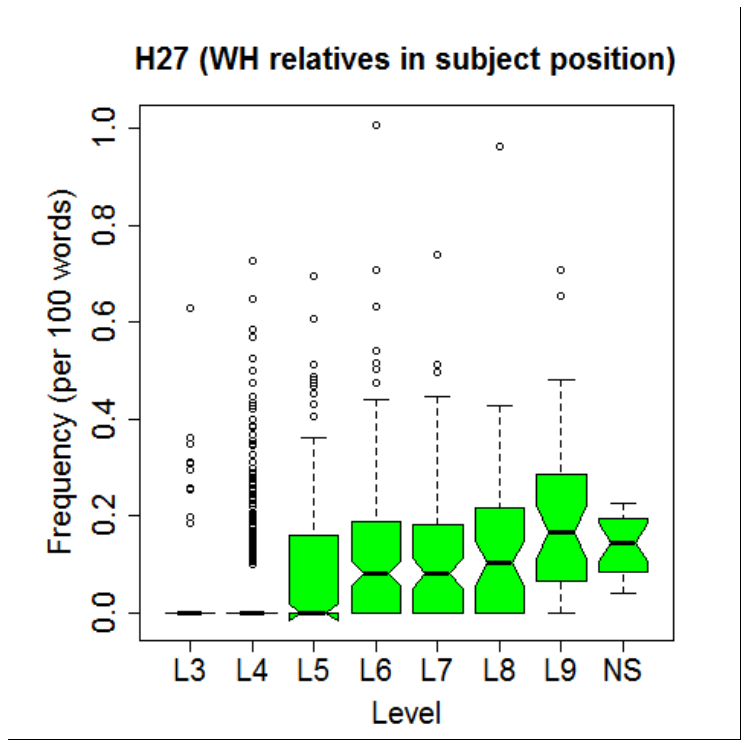


Figure E27. Box-and-whisker plots for H27 (WH relatives in subject position).

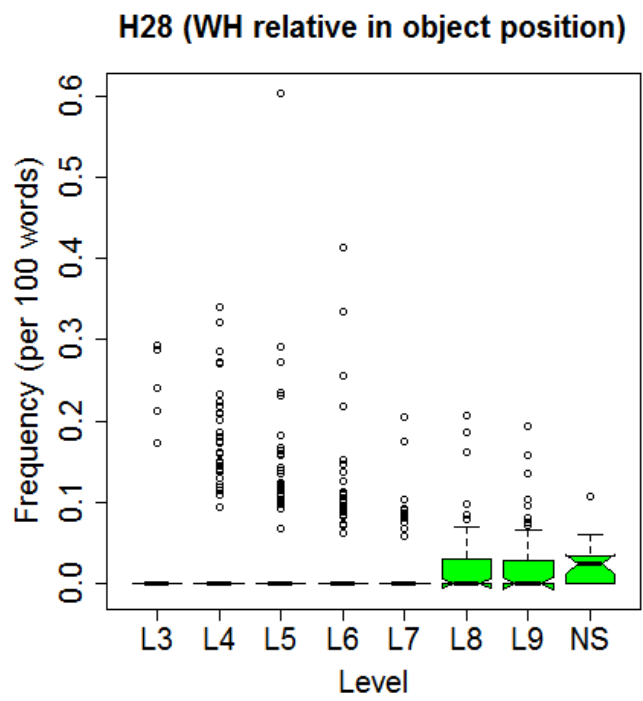


Figure E28. Box-and-whisker plots for H28 (WH relative in object position).

**H29 (WH relatives with fronted preposition)**

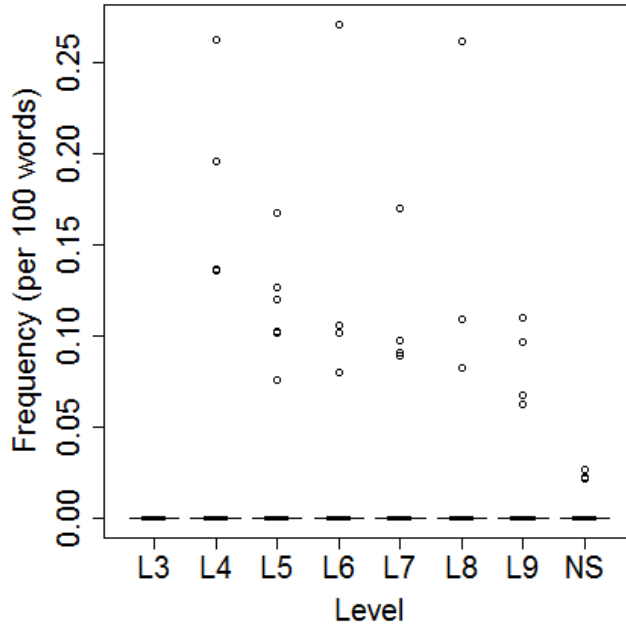


Figure E29. Box-and-whisker plots for H29 (WH relatives with fronted preposition).

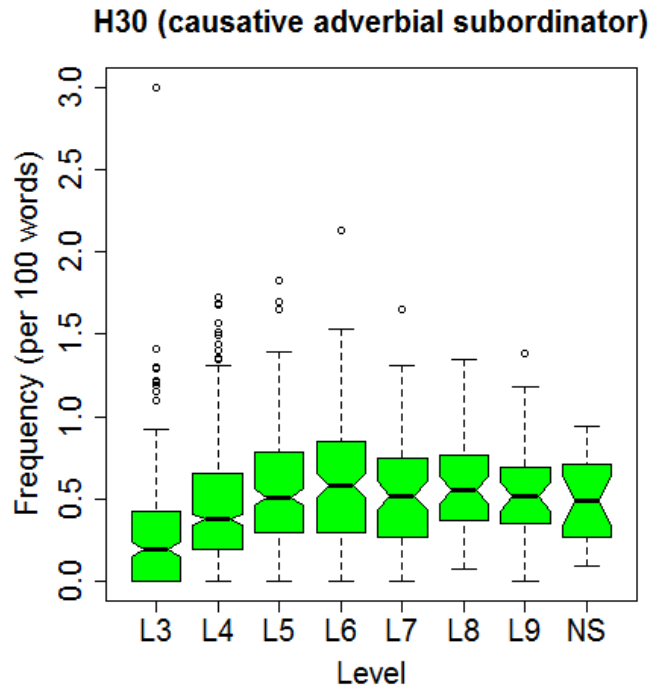


Figure E30. Box-and-whisker plots for H30 (causative adverbial subordinator).

### H31 (concessive adverbial subordinator)

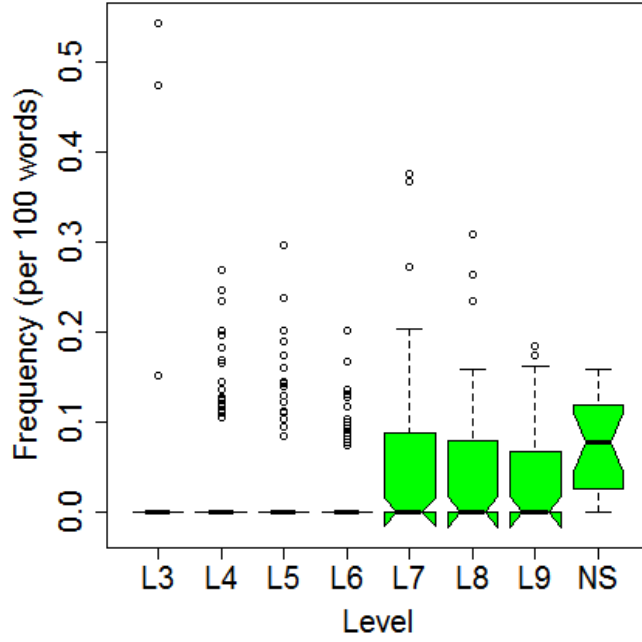


Figure E31. Box-and-whisker plots for H31 (concessive adverbial subordinator).

### H32 (conditional adverbial subordinator)

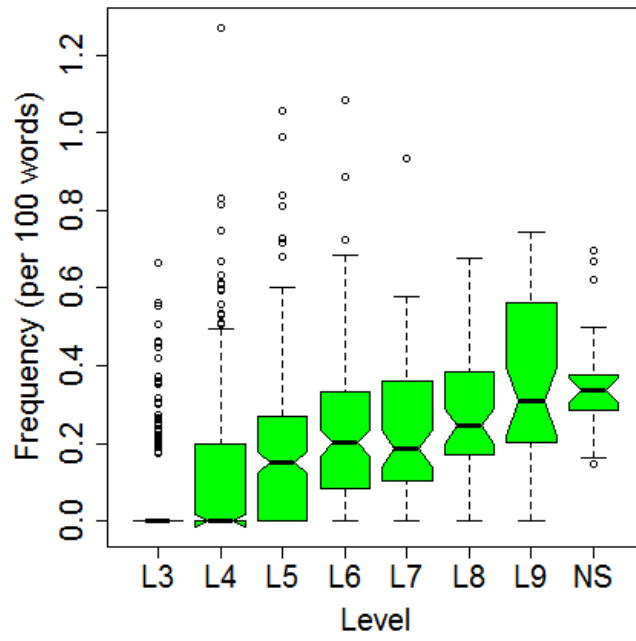


Figure E32. Box-and-whisker plots for H32 (conditional adverbial subordinator).

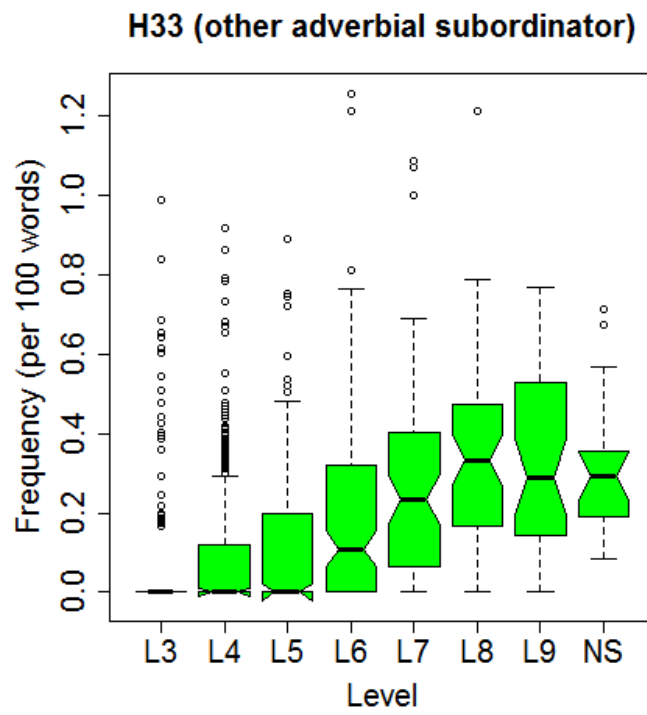


Figure E33. Box-and-whisker plots for H33 (other adverbial subordinator).

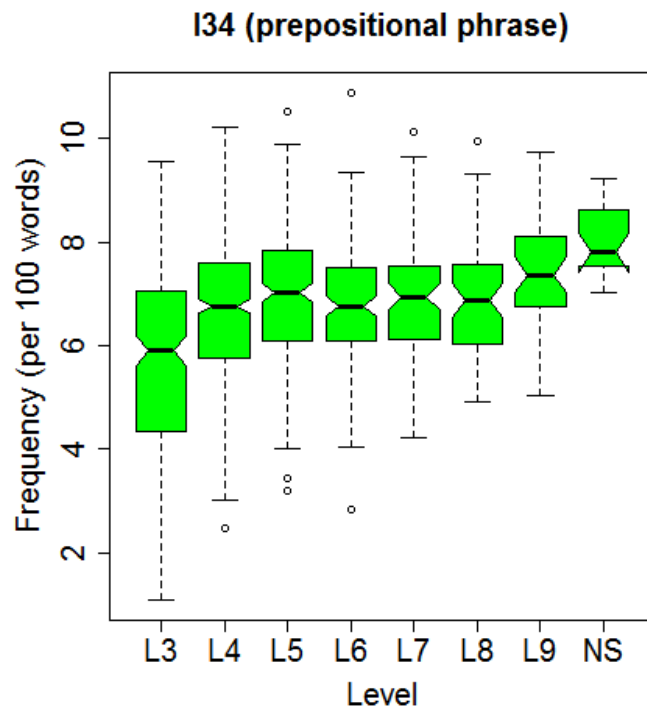


Figure E34. Box-and-whisker plots for I34 (prepositional phrase).

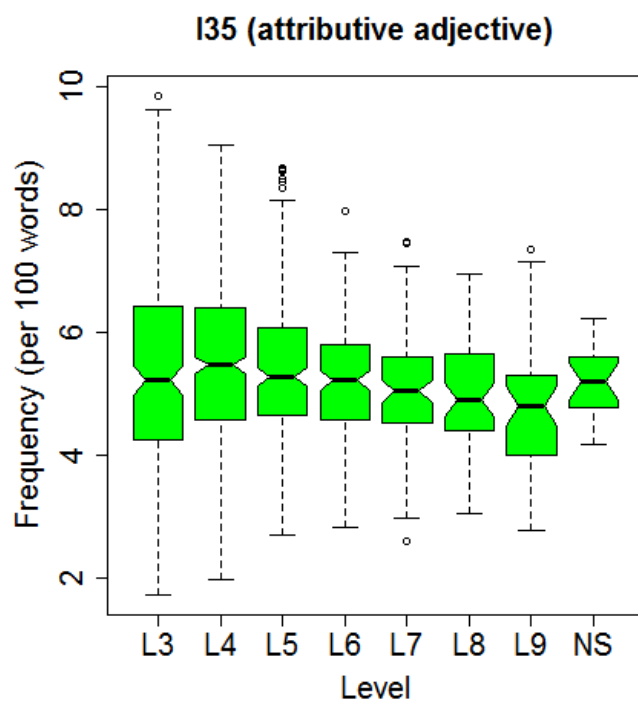


Figure E35. Box-and-whisker plots for I35 (attributive adjective)

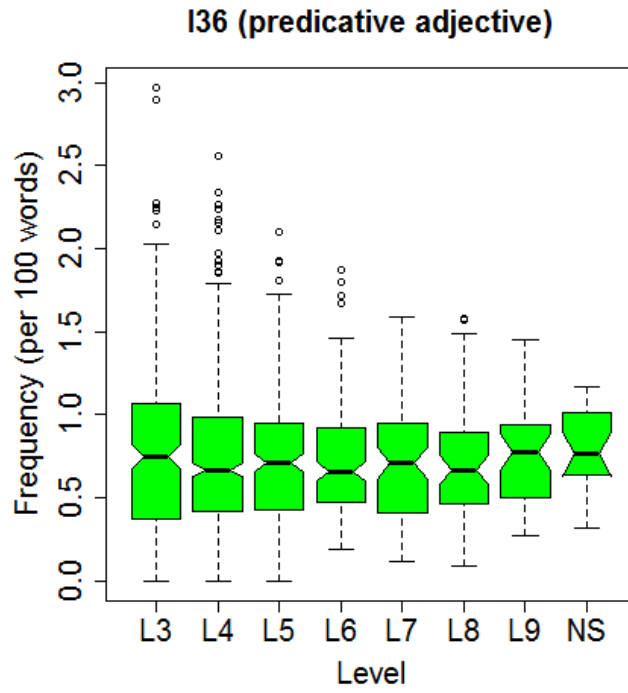


Figure E36. Box-and-whisker plots for I36 (predicative adjective).

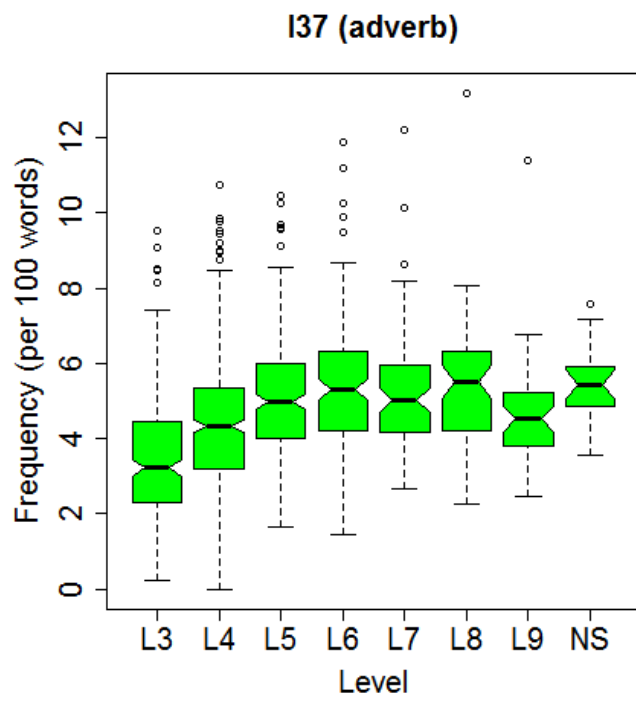


Figure E37. Box-and-whisker plots for I37 (adverb).

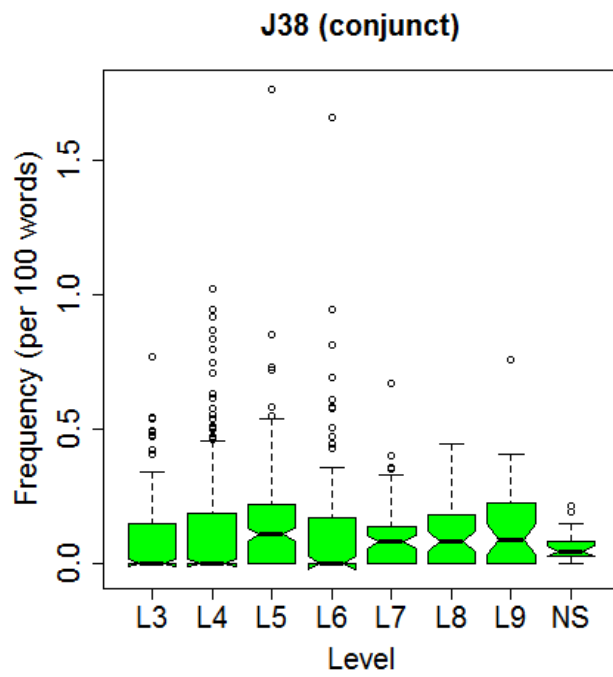


Figure E38. Box-and-whisker plots for J38 (conjunct).

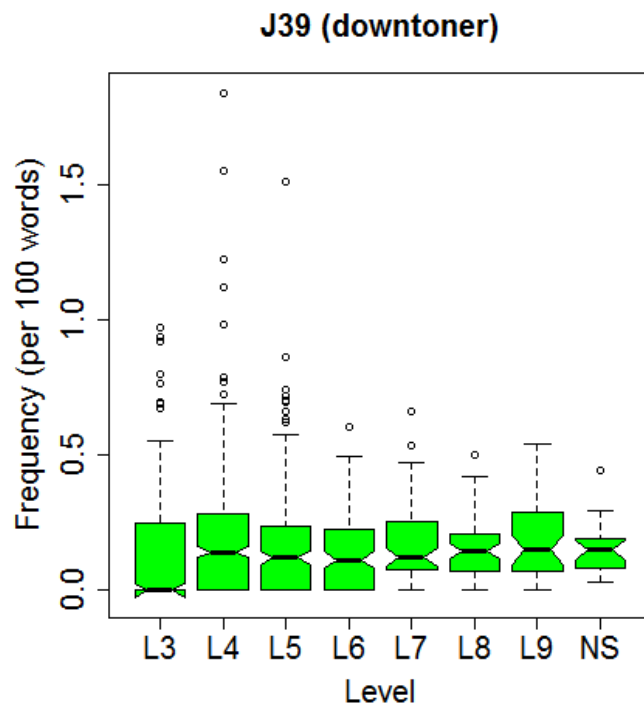


Figure E39. Box-and-whisker plots for J39 (downtoner).

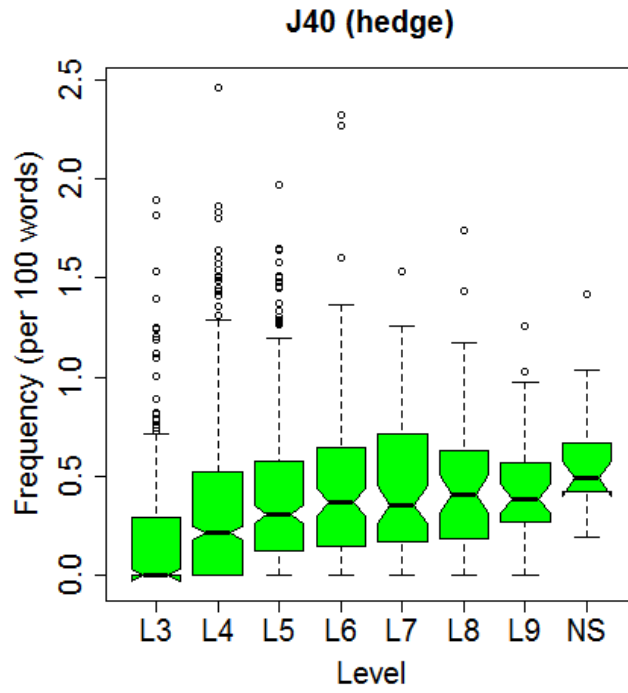


Figure E40. Box-and-whisker plots for J40 (hedge).

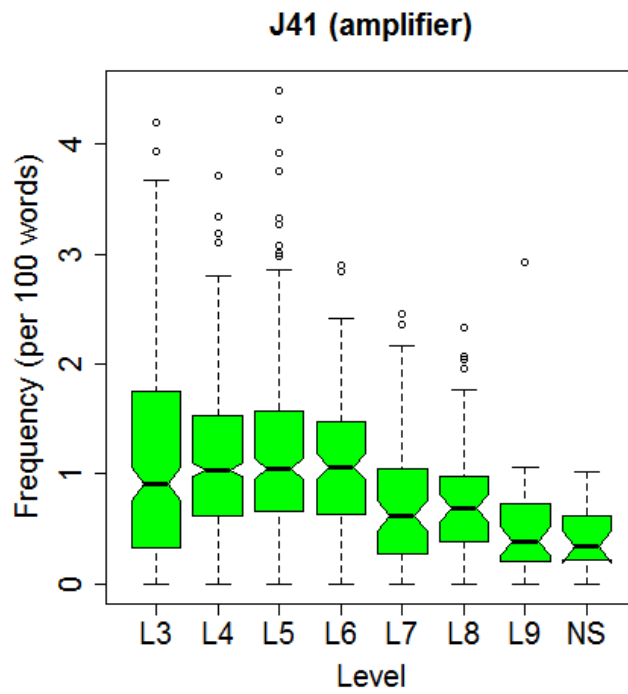


Figure E41. Box-and-whisker plots for J41 (amplifier).

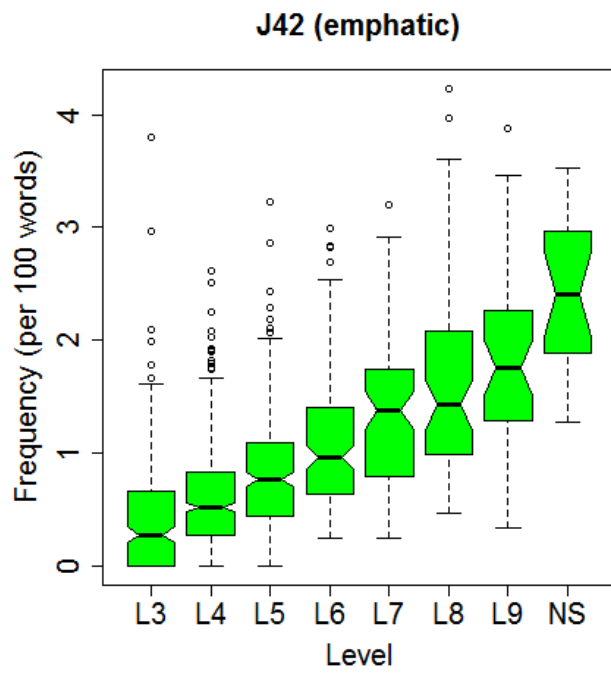


Figure E42. Box-and-whisker plots for J42 (emphatic).

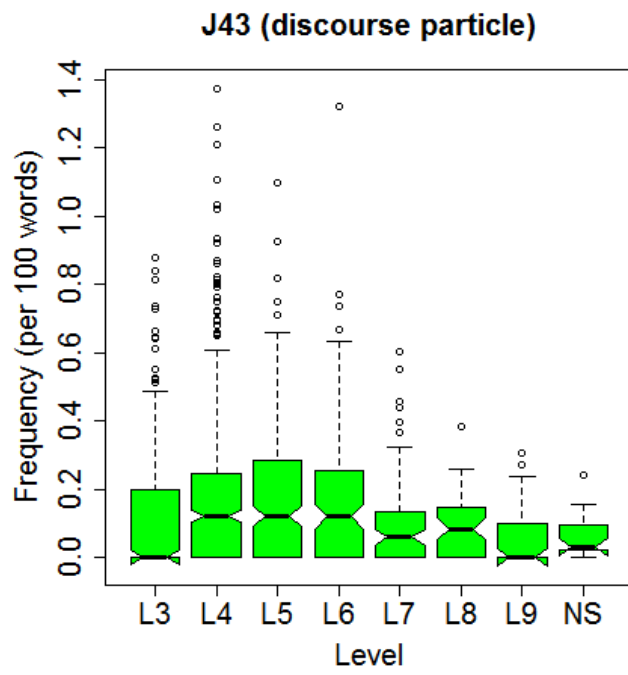


Figure E43. Box-and-whisker plots for J43 (discourse particle).

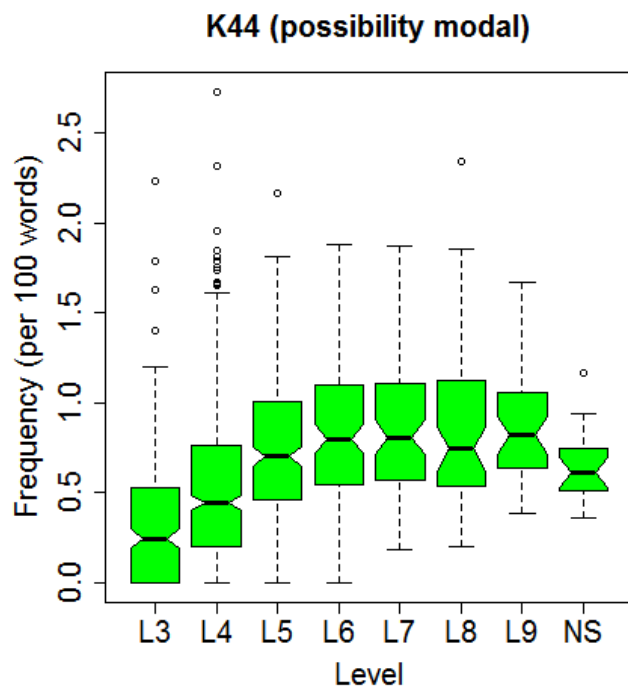


Figure E44. Box-and-whisker plots for K44 (possibility modal).

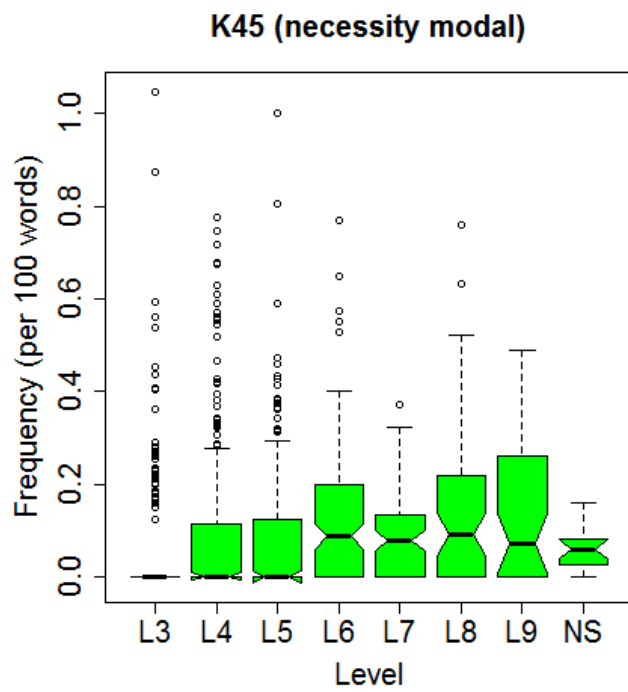


Figure E45. Box-and-whisker plots for K45 (necessity modal).

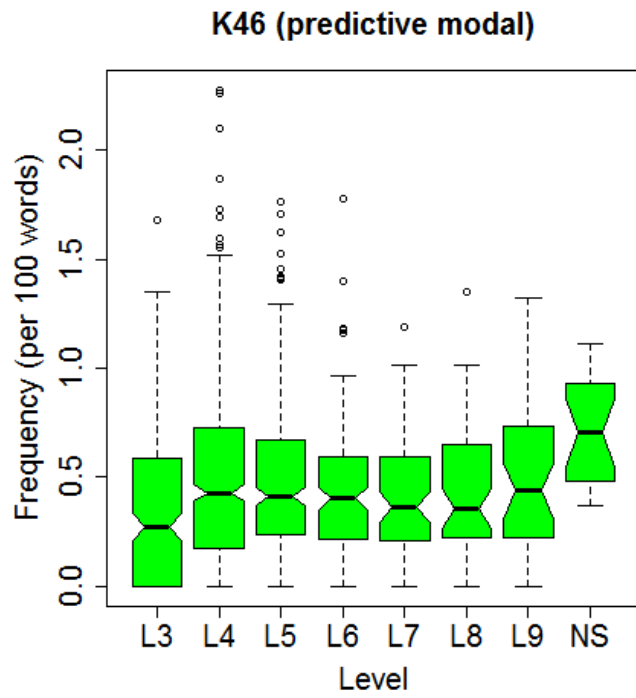


Figure E46. Box-and-whisker plots for K46 (predictive modal).

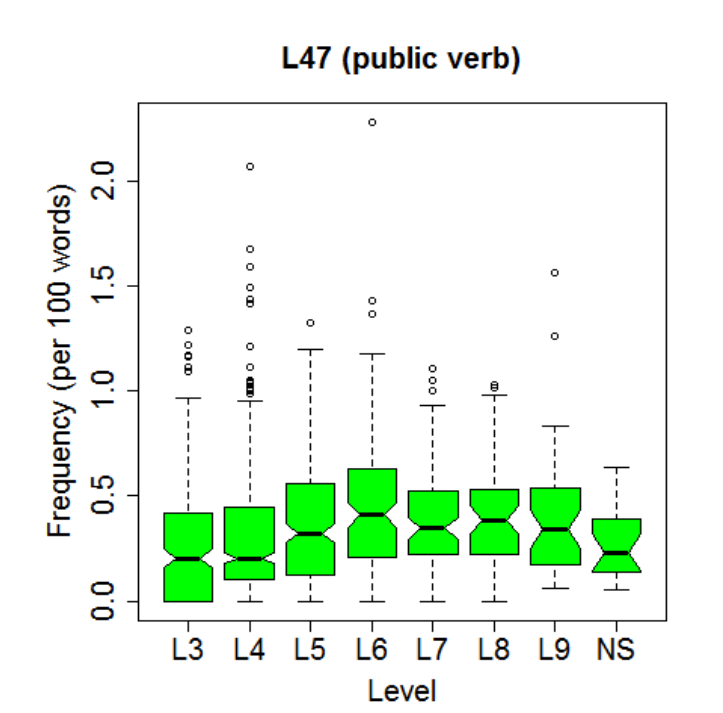


Figure E47. Box-and-whisker plots for L47 (public verb).

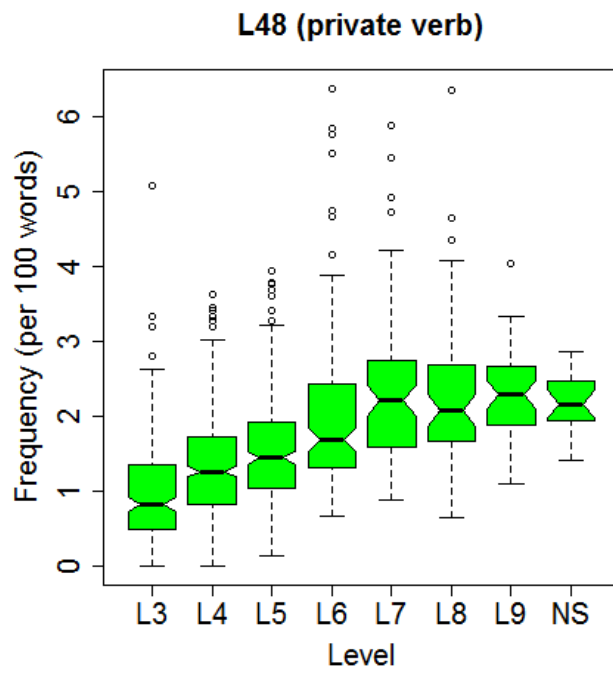


Figure E48. Box-and-whisker plots for L48 (private verb).

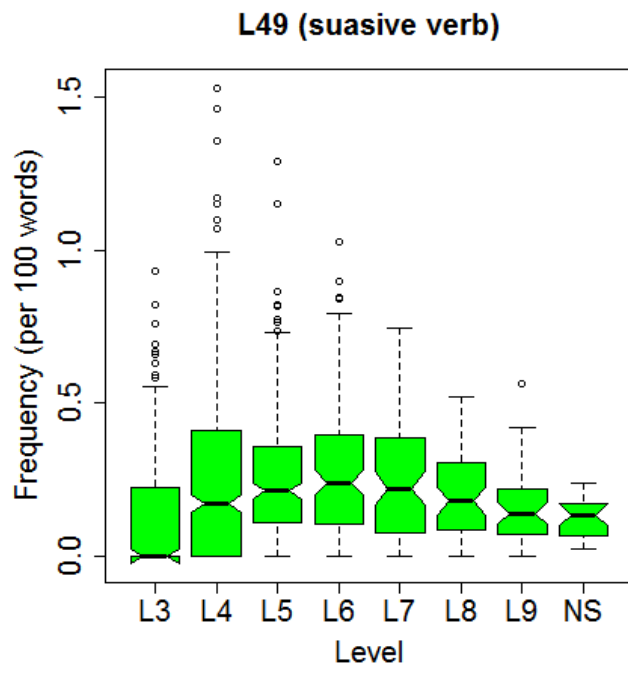


Figure E49. Box-and-whisker plots for L49 (suasive verb).

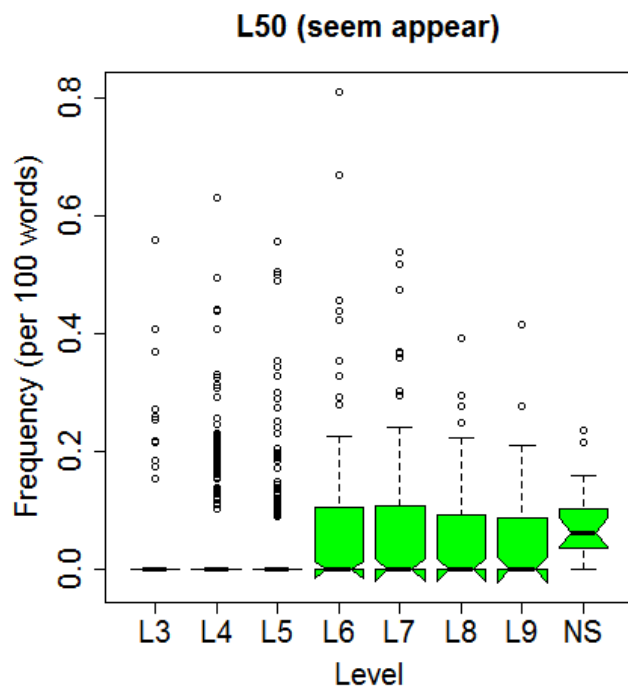


Figure E50. Box-and-whisker plots for L50 (seem appear).

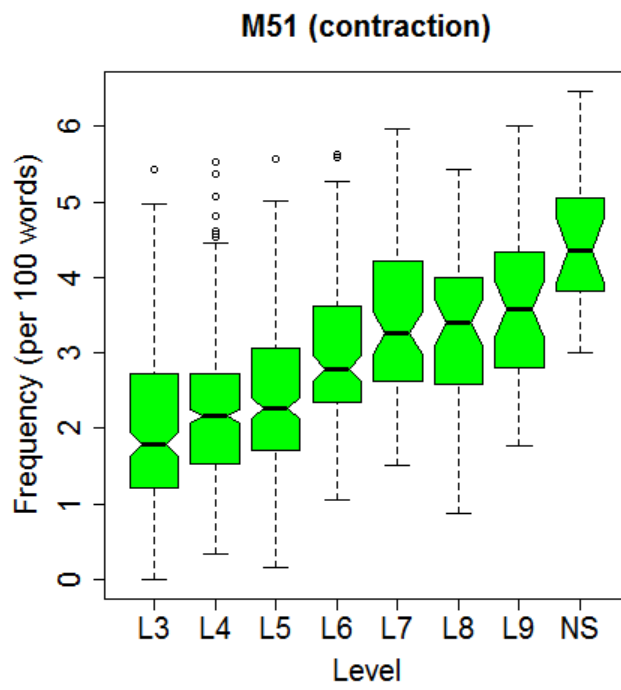


Figure E51. Box-and-whisker plots for M51 (contraction).

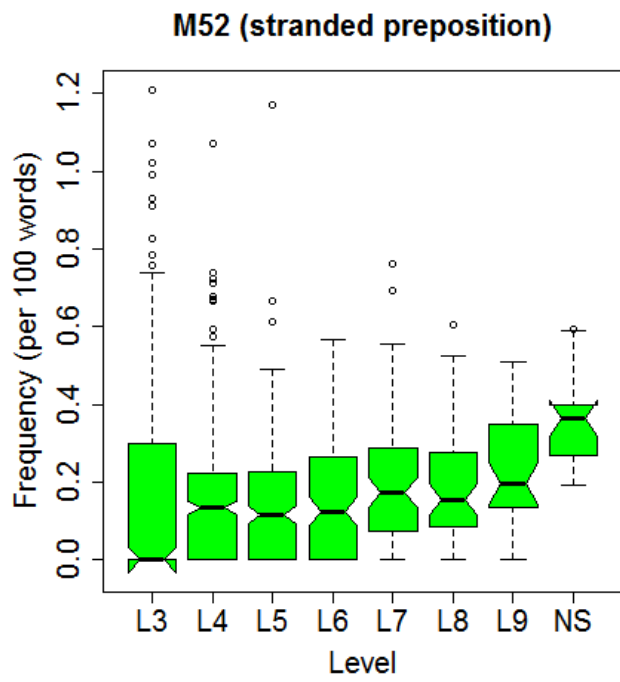


Figure E52. Box-and-whisker plots for M52 (stranded preposition).

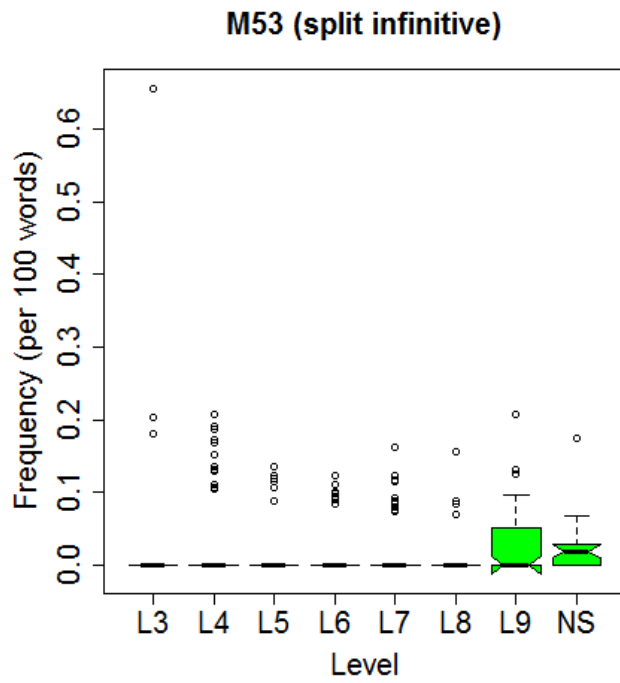


Figure E53. Box-and-whisker plots for M53 (split infinitive).

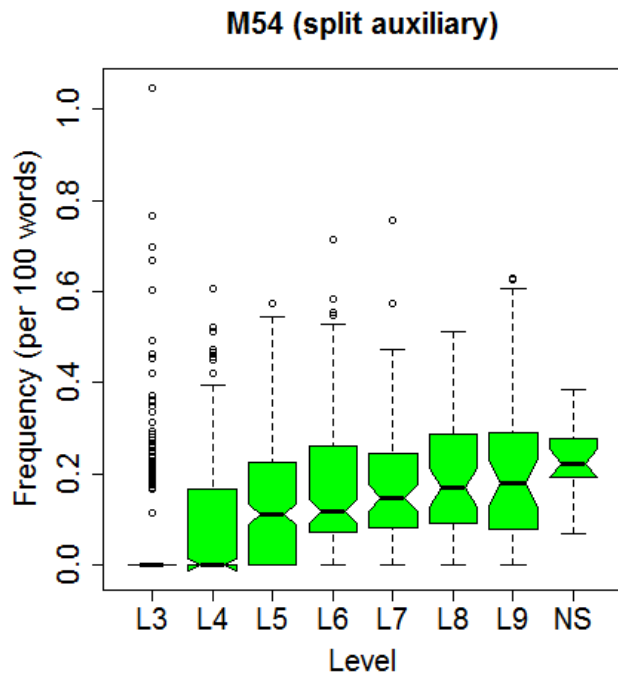


Figure E54. Box-and-whisker plots for M54 (split auxiliary).

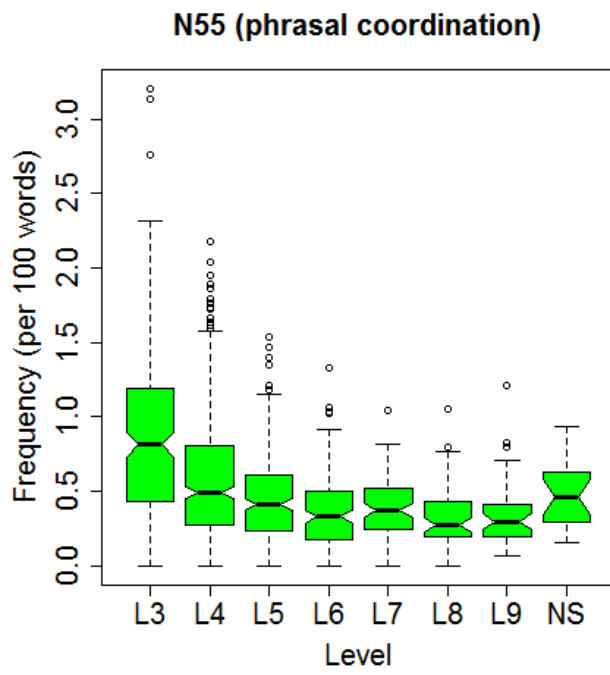


Figure E55. Box-and-whisker plots for N55 (phrasal coordination).

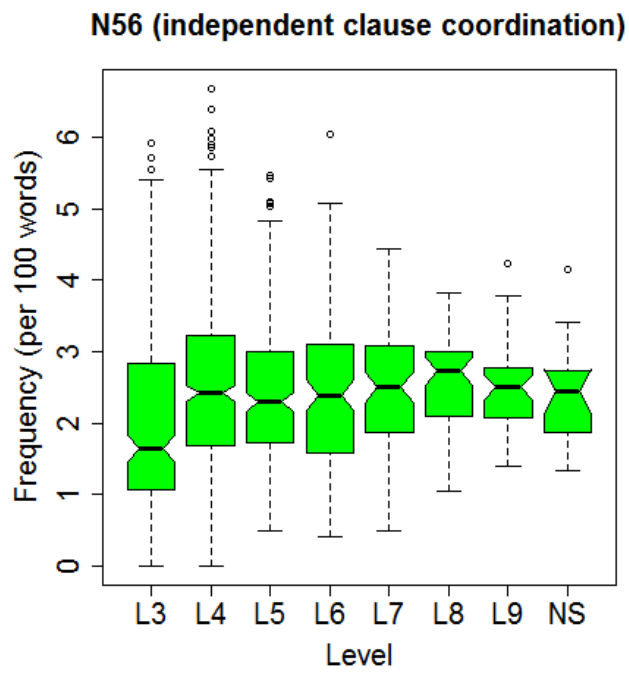


Figure E56. Box-and-whisker plots for N56 (independent clause coordination).

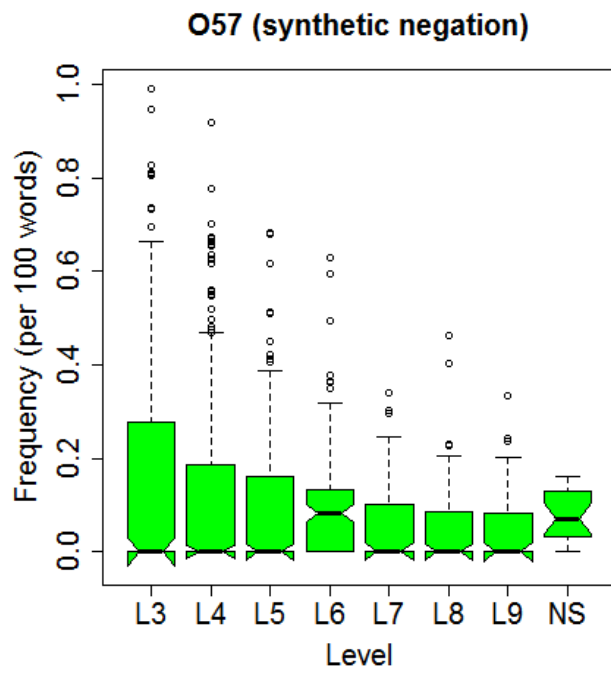


Figure E57. Box-and-whisker plots for O57 (synthetic negation).

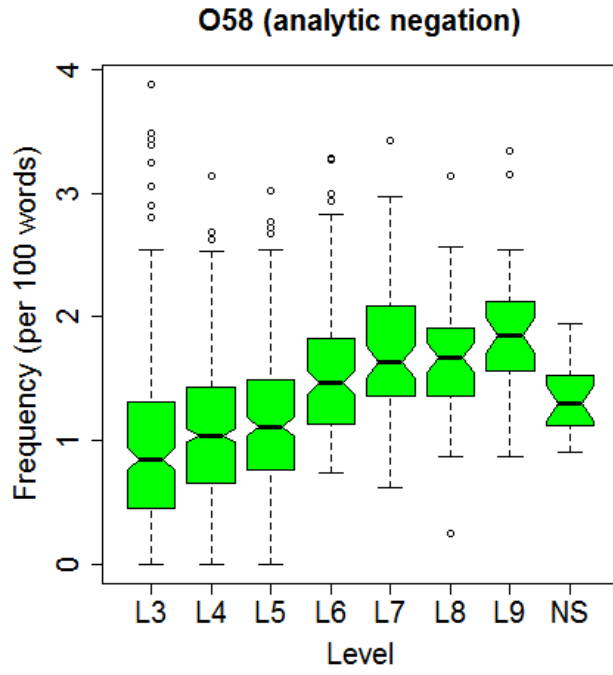


Figure E58. Box-and-whisker plots for O58 (analytic negation).