LINGUISTIC PROFILES OF HIGH PROFICIENCY MANDARIN AND HINDI SECOND LANGUAGE SPEAKERS OF ENGLISH

by

Jie Gao

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of English

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. April Ginther, Chair
Department of English

Dr. Elaine Francis
Department of English

Dr. Tony Silva
Department of English

Dr. Xun Yan
Department of Linguistics, University of Illinois Urbana-Champaign

Approved by:
Dr. Dorsey Armstrong

For mom and dad

Who encourage me to start it off, stick it through, and carry it on

致我的父母

感谢你们给予我的一切

ACKNOWLEDGEMENTS

This dissertation would not have been completed without the support from my mentors, friends, and family, to whom I would like to thank from the bottom of my heart.

My deepest gratitude goes to my advisor, Dr. April Ginther, who has provided a lot of helpful feedback for my dissertation project. The inspiring conversations I had with Dr. Ginther are my most unforgettable experience in graduate school. I would also like to thank Dr. Nancy Kauper, who offered me guidance and mentorship when I worked as a testing office assistant at the Oral English Proficiency Program (OEPP). Those rater training sessions on Friday morning are where this dissertation started!

I am grateful for my dissertation committee members: Dr. Elaine Francis, Dr. Tony Silva, and Dr. Xun Yan. The eye-opening seminars offered by Dr. Elaine Francis and Dr. Tony Silva have enriched my understanding of language studies, and will remain to be great moments in my life. I am deeply thankful for Dr. Xun Yan, whose insightful comments and valuable friendship have been strong support for me during the past few years.

I would also like to thank Dr. Bradley Dilger and Dr. Shelley Staples, who have been great mentors to me. Thanks for involving me in the Corpus and Repository of Writing (Crow) project! I have learned more than I imagine.

I want to thank all of my friends at Purdue: Hadi Banat, Sherri Craig, Ge Lan, Zhi Li, Jingyi Liu, Grace Man, and Carol Chun Zheng. Their company has filled my PhD life with colors and laughter.

Last, but most importantly, I want to thank my parents. Mom and dad, I feel so lucky to be your daughter! Thanks for everything you have given me!

TABLE OF CONTENTS

LIST OF TABLES 7

LIST OF FIGURES 8

ABSTRACT 9

CHAPTER 1. INTRODUCTION 12

CHAPTER 2. LITERATURE REVIEW 19

2.1.Models Explaining L2 Speaking and Differences between L1 and L2 Speech Production 20

2.2.Operationalization of Fluency as the Key Constructs Related with L2 Speaking 24

2.3.Measurement of Utterance Fluency 27

2.3.1.Quantitative Measures of Utterance Fluency 29

2.3.2.....Fluency Variables Selected for Linguistic Profiles of Advanced Intermediate and Advanced L2 English Speakers 33

2.4. Vocabulary—a Dimension in Interaction with Fluency 36

2.4.1.Terminologies used in Vocabulary Studies 38

2.4.2.Lexical Proficiency Development Models 40

2.4.3.References of Vocabulary Breadth Measures 45

2.4.4.Quantitative Measures of Lexical Diversity—a Brief Review 47

2.4.5.Quantitative Measures of Lexical Frequency Profiles 51

2.5. Accentedness Measurement 53

2.5.1.Social Implications of Accentedness 54

2.5.2.A Trinity of Intelligibility, Comprehensibility, and Accentedness 56

2.5.3.Perception of Accentedness 59

2.5.4.Developing Scales to Measure Accentedness 62

2.6. Linguistic profile studies and cluster analysis in applied linguistics research 65

CHAPTER 3. METHODS 69

3.1.Research Question and Design 69

3.2.Database Description 69

3.3.Research Phase I—Variable Coding 71

3.4.Research Phase II—Profile Identification 79

3.5.Research Phase III—Accent Perception 81
CHAPTER 4. RESULTS AND DISCUSSION 88
4.1.Descriptive Statistics and Correlation Results 88
4.2.Principal Component Analysis 93
4.3.Hierarchical Cluster Analysis Results 94
4.4.Sample Selection for Accent Evaluation 104
4.5.Accent Evaluation Results 107
CHAPTER 5. CONCLUSION AND IMPLICATION 120
5.1.Profile Information of Advanced Intermediate and Advanced L2 English Speakers 120
5.2.Connections among L2 speakers' Overall L2 Proficiency, L1 Background, Accentedness, Fluency, and Vocabulary Features 122
5.3.The distribution of Advanced Intermediate and Advanced L2 English Speakers in the Four Clusters 125
5.4.Implication for Rating and Future Pedagogy Design 125
5.5.Limitation of the Study 127
APPENDIX A 129
APPENDIX B 130
APPENDIX C 131
APPENDIX D 135
APPENDIX E 137
APPENDIX F 145
REFERENCES 152
VITA 180

LIST OF TABLES

Table 2.1 Strength of Accent Scale (Ockey & French, 2016) 64
Table 3.1 Transcribed Speech Samples from OEPT 2 71
Table 3.2 Measured Constructs and Coded Variables 73
Table 3.3 Special Characters Used for Transcribing Speech by Fluencing 76
Table 3.4 Rater Assignment for Speech Accent Evaluation 84
Table 3.5 Adapted Accentedness Measurement Scale from Ockey and French (2016) 85
Table 3.6 Excerpt Data Collection for Accent Evaluation 87
Table 4.1 Descriptive Statistics of Fluency and Vocabulary Measures 88
Table 4.2 Correlation between Variables for All Speakers 92
Table 4.3 Component Loadings for Speakers Rated as 50 (after Promax Rotation) 94
Table 4.4 Four Levels to Describe the Five Features across the Two Proficiency Levels 96
Table 4.5 Hypothesized Profile Information after Hierarchical Cluster Analysis (HCA) 97
Table 4.6 Hypothesized Sample Selection for Accent Evaluation 97
Table 4.7 Descriptive Statistics of Component 1 (Fluency Features) Score and Component 2 (Vocabulary Features) Score across Clusters 98
Table 4.8 Ordinal Scale Conversion of Fluency and Vocabulary Features 98
Table 4.9 Descriptive Statistics of Fluency and Vocabulary Measures across Clusters 99
Table 4.10 Description of Mean Fluency and Vocabulary Measures for Each Cluster 100
Table 4.11 Cluster Membership Information 101
Table 4.12 Speech Sample Selection Chart 105
Table 4.13 Information of Selected Speech Samples 106
Table 4.14 Descriptive Statistics of Accent Evaluation Results 107
Table 4.15 Ordinal Scale for Accent Evaluation 116
Table 4.16 Logit Scale for L1 Hindi Speakers 116
Table 4.17 Logit Scale for L1 Mandarin Speakers 117
Table 4.18 Profile Information for all the Speakers across Cluster 119

LIST OF FIGURES

Figure 2.1 A blueprint for the speakers (Levelt, 1989:9) 20
Figure 2.2 A blueprint of the speaker (Levelt, 1999: 68) 21
Figure 2.3 The Model of Bilingual Speech Production (Kormos, 2006: 168) 23
Figure 2.4 The lexical space: dimensions of word knowledge and ability (Daller, Milton & Treffers-Daller, 2007: 8) 43
Figure 2.5 The Cube (from You, 2014: 15) 44
Figure 2.6 Ideal TTR versus Token Curves (from Malvern et al., 2004: 52) 49
Figure 3.1 Sample Screenshot of the Fluencing Annotation Tool (Park, 2016: 46) 75
Figure 3.2 Fluencing Output of Temporal Variables 77
Figure 3.3 Interface of AntWord Profiler 78
Figure 3.4 Vocabulary Frequency Information Retrieved from AntConc Word Profiler 78
Figure 3.5 Principal Component Analysis 80
Figure 3.6 MFRM Model Equation (Eckes, 2011:14) 86
Figure 4.1 Boxplots of Fluency and Vocabulary Measures across Proficiency Levels 89
Figure 4.2 Scree Plot for Principal Component Analysis of Fluency and Vocabulary Measures 93
Figure 4.3 Hierarchical Cluster Analysis Dendrogram 95
Figure 4.4 Hierarchical Cluster Analysis Scree Plot 95
Figure 4.5 Cluster Membership of L1 Hindi Speakers Rated as 50 102
Figure 4.6 Cluster Membership of L1 Mandarin Speakers Rated as 50 103
Figure 4.7 Cluster membership of L1 Hindi speakers rated as 60 104
Figure 4.8 Cluster membership of L1 Mandarin speakers rated as 60 104
Figure 4.9 Accent Evaluation Results for Cluster 1 109
Figure 4.10 Accent Evaluation Results for Cluster 3 111
Figure 4.11 Accent Evaluation Results for Cluster 2 112
Figure 4.12 Accent Evaluation Results for Cluster 4 114
Figure 4.13 Wright Map for MFRM Results 115

ABSTRACT

This dissertation investigates three utterance fluency features and two vocabulary features of 409 speech samples from advanced intermediate and advanced L2 English speakers, who participated in the Oral English Proficiency Test (OEPT) between the year of 2009 and 2015. Among the 409 L2 English speakers, there are 80 L1 Hindi speakers rated as advanced intermediate, 32 L1 Hindi speakers rated as advanced, 286 L1 Mandarin speakers rated as advanced intermediate, and 11 L1 Mandarin speakers rated as advanced.

Hierarchical Cluster Analysis (HCA) was conducted and presented four different clusters among all the L2 English speakers. The four different clusters are: (1) Low Mean Syllables per Run (MSR), low Speech Rate (SR), very high Pause Rate (PR), medium Measure of Textual Lexical Diversity (MTLD), and medium percentage of words on the Academic Word List (AWL); (2) Medium Mean Syllables per Run (MSR), medium Speech Rate (SR), high Pause Rate (PR), low Measure of Textual Lexical Diversity (MTLD), and low percentage of words on the Academic Word List (AWL); (3) High Mean Syllables per Run (MSR), high Speech Rate (SR), low Pause Rate (PR), medium Measure of Textual Lexical Diversity (MTLD), and medium percentage of words on the Academic Word List (AWL); (4) Medium Mean Syllables per Run (MSR), medium Speech Rate (SR), low Pause Rate (PR), very high Measure of Textual Lexical Diversity, and very high percentage level of words on the Academic Word List (AWL).

Chi-square results show that L2 English speakers' cluster membership is strongly associated with both their L1 background and level of L2 oral English proficiency. While most of the advanced intermediate L1 Mandarin speakers are in Cluster 1 and Cluster 2, the majority

of the advanced intermediate L1 Hindi speakers concentrate in Cluster 3. A large number of advanced L1 Mandarin speakers and L1 Hindi speakers are also located in Cluster 3.

Twelve raters were invited to evaluate speech samples representative of the four clusters in terms of accent difference and listener effort. Twelve speakers were selected from the four clusters, whose speech samples have values of the five linguistic features closest to the cluster mean.

Multi-facet Rasch Measurement (MFRM) results show that L1 Mandarin speakers generally received lower ratings in accent difference and listener effort. The connection among fluency, vocabulary, and accentedness/listener effort, however, functions differently for L1 Mandarin speakers and L1 Hindi speakers. For advanced intermediate L1 Mandarin speakers, those who speak slower and use more diverse vocabulary and more academic words were evaluated to be less accented, meanwhile costing less listener effort. However, advanced intermediate L1 Hindi speakers were rated as less accented and cost less listener effort when they demonstrate higher fluency measures and lower vocabulary measures.

Advanced L2 English speakers, in contrary, received reverse rating results. The advanced L1 Mandarin speaker, who speaks faster and uses less diverse vocabulary and fewer academic words, was evaluated to be less accented and cost less listener effort. However, the advanced L1 Hindi speaker, who speaks slower and uses more diverse vocabulary and more academic words, was rated as less accented and cost less listener effort.

This dissertation reemphasizes that holistic rating rubric does not deny the existence of multiple linguistic profiles. Raters are sensitive to different combinations of fluency and vocabulary features even if they have been asked to use a holistic scale. In addition, L2 English

speakers may adopt individual strategies to accommodate while delivering, which calls for further pedagogical attention.

Key words: Cluster Analysis, Fluency, L2 Speaking, Linguistic Profiles, Rater Interaction, Vocabulary

CHAPTER 1. INTRODUCTION

From the perspective of testing and assessment, speaking proficiency can be represented by a number of components, typically fluency, grammar, vocabulary, and pronunciation. During the holistic rating process of an oral language proficiency test, these factors, and others, contribute to an examinee's final score in combination. This dissertation investigates aspects of fluency, vocabulary, and pronunciation for examinees at higher language proficiency levels on the Oral English Proficiency Test (OEPT), Purdue's oral English assessment for prospective international teaching assistants (ITAs). Linguistic profiles focusing on fluency and vocabulary will be established through a cluster analysis of objective measurement indices for advanced intermediate and advanced (OEPT scores of 50 and 60) L2 English speakers, whose first language background is Hindi or Mandarin.

After obtaining linguistic profiles through cluster analysis, trained OEPT raters were asked to evaluate accentedness of test takers within each profile. The study synthesizes and extracts selected components of language proficiency in order to establish linguistic profiles for examinees of different L1s, who are located at the higher levels of a scale but typically have very different accents. Examinee performance rated with the same score might demonstrate different or similar characteristics in fluency and vocabulary, and these in turn may simultaneously influence raters' perceptions of accentedness. Also, the growth of second language proficiency from advanced intermediate to advanced may display different combinations among the three components. Examining how clusters of abilities may emerge enriches our understanding of the effects of L1 backgrounds and proficiency levels on speaking performance.

OEPT examinees have diverse language and cultural backgrounds. L1 Mandarin Chinese and L1 Indian Hindi speakers constitute the two largest groups of test takers. The responses of examinees across these L2 groups differ considerably with respect to grammar, vocabulary, and pronunciation. Examinees from India tend to score higher than examinees from China; however, examinees from both groups overlap at the advanced intermediate level. Gathering related profile information strengthens the connections between test performance and the holistic rating scale. While final scores place examinees within levels in terms of language proficiency, achieving the same score does not indicate absolute homogeneity or an elimination of performance diversity. In addition, differences in linguistic performance of L1 Hindi and L1 Mandarin speakers can be attributed to educational contexts and language learning experience. English has been used as an instructional language in Indian educational institutes and is extensively spoken as a second language. Most students with an L1 Mandarin background, however, learn English as a foreign language and usually pay a large amount of attention to language test preparation.

A better understanding of performance profiles sheds light on the performance of higher-level L2 speakers. In comparison to test responses rated as the lowest level or the highest level, those scored as medium high have not received the same amount of attention. Advanced intermediate L2 English speakers, who have fulfilled university admission requirement for language proficiency, are often exempted from post-admission English tests or language courses. Description with this profile information of advanced intermediate and advanced L2 English speakers would benefit from more detailed explication of fluency variables, along with the co-functioning of lexical complexity and other potentially influential factors, such as speech accentedness.

To conduct a descriptive analysis of linguistic profiles displayed by OEPT English L2 speakers, this dissertation partially relies on the measurement of complexity, accuracy, and fluency (CAF) as framework, meanwhile exploring the application of CAF to second language speaking. Complexity, accuracy and fluency (CAF) have been employed in order to address the multi-componentiality of L2 proficiency. These dimensions are simultaneously recognized as goals for language task performance as well as research variables (Skehan, 1996, 1998; Housen, Kuiken, & Vedder, 2012). Related research is initiated on the ground of providing clear operationalizations for the three dimensions, selecting subcomponent measurements and seeking for operationalization methods.

Complexity, fluency, and accuracy (CAF) features have been primarily applied to studies of writing. For instance, Wolfe-Quintero, Inagaki, and Kim (1998) includes a research outline in the form of a technical report, which examined second language writing development with the CAF framework. Language production units such as clauses, T-units, and sentences are common measures for the three dimensions. Fluency in writing is represented by the frequency and length of production units, among which error-free ones are accepted as demonstrations of accuracy. Complexity constitutes of grammatical complexity and lexical complexity. In Wolfe-Quintero, Inagaki, and Kim (1998), measurement of fluency and accuracy depends on the counts and indices of various clause types or grammatical structures, while complexity is represented by lexical variation, density, and sophistication with type-token ratio calculations.

In the assessment of second language speaking, further subconstruct analysis and interpretation is provided by Foster, Tonkyn, and Wigglesworth (2000), who identify the analysis of speech unit (AS-unit) as a main syntactic unit for spoken language research: “An AS-unit is a single speaker’s utterance consisting of an independent clause, or sub-clause unit,

together with any subordinate clause(s) associated with either" (p. 365). Their focus on micro-level units helps keep track of the relationships between complexity, accuracy, and fluency in L2 speaking.

More rigorously defined language production units have facilitated various methods for subconstruct measurement in second language acquisition, which contain but are not confined within: (a) the definition of the three constructs or investigation of a specific domain, explaining dynamics and connections among the three categories under the roof of CAF. Research studies developed from CAF definitions involve discussion of CAF operationalizations, such as the explanation of syntactic complexity together with clause-based and length-based metrics (Norris & Ortega, 2009), a more comprehensive and accurate taxonomic model of L2 complexity (Bulté & Housen, 2012), or a definition clarification of CAF constructs and caution for operationalization (Pallotti, 2009); (b) the application of both global and specific local measures for weighing complexity, accuracy and fluency (Larson-Freeman, 2006; Tonkyn, 2012); (c) task properties and their influence on linguistic performance, which is represented by syntactic complexity, types of speech fluency, and lexical diversity. Correlations among the three dimensions (e.g. the correlation coefficient between complexity and fluency measures in both L2 speaking and L2 writing) are usually explained by trade-off effects and models of cognitive account (De Jong et al., 2012a; Kuiken & Vedder, 2012; Levelt, 1989, 1999a; Robinson, 2001c; Révész, Ekiert, & Torgersen, 2016; Skehan, 2009b).

Complexity, accuracy, and fluency (CAF) have been conceptualized from both macro and micro perspectives. Skehan (2003) describes the three dimensions globally in speaking with phrases such as "greater control of the emerging system" and "new interlanguage elements are used not simply haltingly and incorrectly" (p. 8), but also lists studies investigating specific

measures targeting at complexity and accuracy. In Housen, Kuiken, and Vedder (2012:3), where researchers review complexity, accuracy, and fluency of overall L2 language proficiency development, they generalize complexity, accuracy, and fluency as "internalization of new L2 elements", "modification of L2 knowledge", and "consolidation and proceduralization of L2 knowledge" accordingly, which is more of a bird eye review in analyzing the "multilayered, multifaceted, and multidimensional" CAF constructs. Whether more specific measures are needed for L2 speaking assessment is an interesting question. Specific measures of complexity, accuracy, and fluency (CAF) have been extensively used to investigate the validity of L2 English speaking test (Yan, Kim, & Kim, 2018), or evaluate second language speakers' progress within study abroad contexts (Juan-Garau, 2014, 2018; Pérez-Vidal & Juan-Garau, 2011; Valls-Ferrer & Mora, 2014).

This dissertation is grounded on a fine-grained interpretation of CAF as applied to speaking performance. First, fluency is represented by separate fluency variables that are usually applied to speaking assessment, which are strongly correlated with language proficiency and differ speakers across proficiency levels. However, investigating speech within a proficiency level, especially that of higher proficiency levels, also necessitates examination of both condensed and extended dimensions of fluency. In addition, major research questions were formulated based on vocabulary usage. Vocabulary factors, such as lexical complexity, have been categorized as either complexity or a separate dimension in addition to CAF (Wolfe-Quintero, Inagaki, & Kim, 1998; Skehan, 2009b).

In this dissertation, accentedness is another important component in extracting L2 English speakers' linguistic profile. As a construct that can only be evaluated in speaking performance, accentedness might have broader interpretation than being against pronunciation

accuracy or deviating from inner-circle English language norms. Derwing and Munro (2009) argue that being accented does not necessarily lead to communication difficulties – i.e., problems with intelligibility and comprehensibility. In contrast, even low levels of perceived accentedness has been identified to negatively influence listeners' performance on listening comprehension tasks (Ockey & French, 2016; Ockey, Papageorgiou, & French, 2016). For higher-level L2 English speakers, however, accentedness is neither a solid representation of pronunciation accuracy nor in a trade-off relationship with fluency. Speakers who are rated at the highest levels of a proficiency scale can still have identifiable pronunciation patterns distinctive from those of L1 English speakers, as the dimension of accentedness may not always observe a "progressive" development pattern along with overall language proficiency. A discussion of accentedness helps broach new research questions for this dissertation: Could trained raters' perception of accentedness be influenced by fluency and lexical factors? Could higher-level L2 speakers experience any change in accentedness as their overall language proficiency grows?

I hypothesized that examinees who speak with faster speech rate and use vocabulary of greater complexity may also sound more accented to the listeners. L1 Hindi speakers might experience a drop in accentedness evaluation when their L2 English proficiency grows from advanced intermediate to advanced. For L1 Mandarin speakers, however, the major change would happen to their use of vocabulary. To be more specific, L1 Mandarin speakers with higher L2 English proficiency may use more diverse vocabulary, but would be rated in a similar manner in terms of accentedness.

Three phases of study are designed with measurement of accentedness built in as the last step. The first phase is a quantification process, which involves a discussion of common measurement indices in evaluating L2 English oral proficiency. Temporal fluency and lexical

usage are examined to create linguistic profiles of speakers at advanced intermediate and advanced levels. The second phase includes a cluster analysis, which presents-detailed classification information about emergent profiles. The last phase focuses on rater perception, where trained raters of OEPT are asked to evaluate accentedness. This dissertation will:

a) Identify operationalizable variables and develop linguistic profiles for L2 English speakers who were rated as advanced intermediate and advanced in OEPT.

b) Investigate raters' perception of accentedness for each profile and possible influence introduced by fluency and lexical factors. Raters' perception of accentedness may vary based on different combinations of fluency and vocabulary features. In this study, raters were also asked to evaluate speakers' accent in terms of the test taker's difference from the local norm and to rate possible comprehensibility difficulties that may be caused. That is, raters may spend a greater amount of effort to concentrate while listening to speech they identify as displaying greater differences from local norms.

CHAPTER 2. LITERATURE REVIEW

Before selecting key dimensions to establish linguistic profiles for second language English speakers, a critical step is to examine the commonly used approaches for understanding speaking. In applied linguistics and language testing, four research themes permeate the discussion of the characteristics, development, and assessment of speaking proficiency: (a) models explaining second language (L2) speaking and differences between first language (L1) and L2 speech production; (b) operationalization of the key constructs related with L2 speaking; (c) the place and importance of pronunciation and accent, and (d) the shift from the focus on pronunciation and accentedness to the now familiar trinity of accentedness, comprehensibility, and intelligibility. In this chapter of literature review, I start with a cognitive approach in understanding speaking, and select fluency and vocabulary as representative constructs in speaking assessment. The discussion about operationalizable constructs and pronunciation related topics reflects listeners' perception, whose participation and reaction are of paramount importance in speaking assessment.

While fluency, lexical factors, and accentedness are the key dimensions discussed in literature review, this chapter also presents an inventory for quantitative measures of utterance fluency and mathematical calculations of lexical diversity. The purpose of building an inventory that documents related indices used in previous studies, as mentioned in Segalowitz, French and Guay (2017), is "providing advance guidance as to which features to look at, thereby avoiding 'fishing expeditions' to find appropriate features on which to focus."(p. 105). A catalog of variables with sufficient explanation and background information renders researchers with a tool kit, or lists of reference at least, to represent speaking-related constructs.

2.1. Models Explaining L2 Speaking and Differences between L1 and L2 Speech Production

Among the explanatory models providing explanation for speech production, the widely-cited and adapted modular model proposed by Levelt (1989) has played a vital role. The model identifies three major components of speech production: the conceptualizer, formulator, and articulator. Reliant on the lexicon as a knowledge store, the three components support the speaking process in concert. Speakers, acting as information processors, are argued to deploy two major conceptual generation processes in speech production: macroplanning for retrieving message, and microplanning for providing new information.

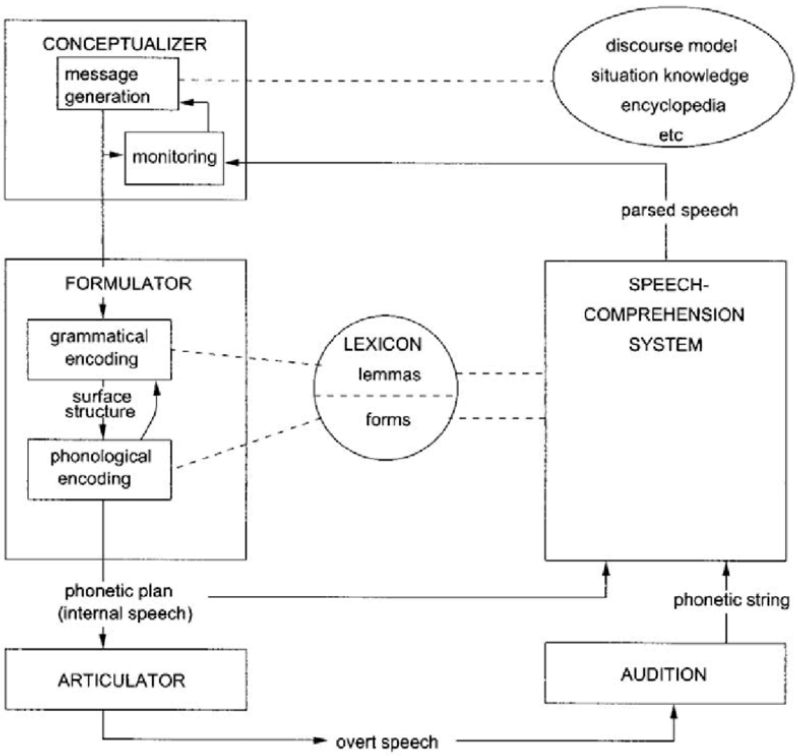


Figure 2.1 A blueprint for the speakers (Levelt, 1989:9)

In the Blueprint Model, a later adaptation by Levelt (1999a), speaking begins with a conceptual preparation stage, passes on to the phase of grammatical, morpho-phonological and phonetic encoding, and is finally realized by articulating overt speech. Lemmas, which are syntactic words in the mental lexicon, are activated at the earliest phase. Lemma activation establishes the surface structure of language output with syntactic structures, and further triggers morpho-phonological encoding. As the last step immediately preceding overt speech production, phonetic encoding results in the ultimate articulation of syllabary outcomes.

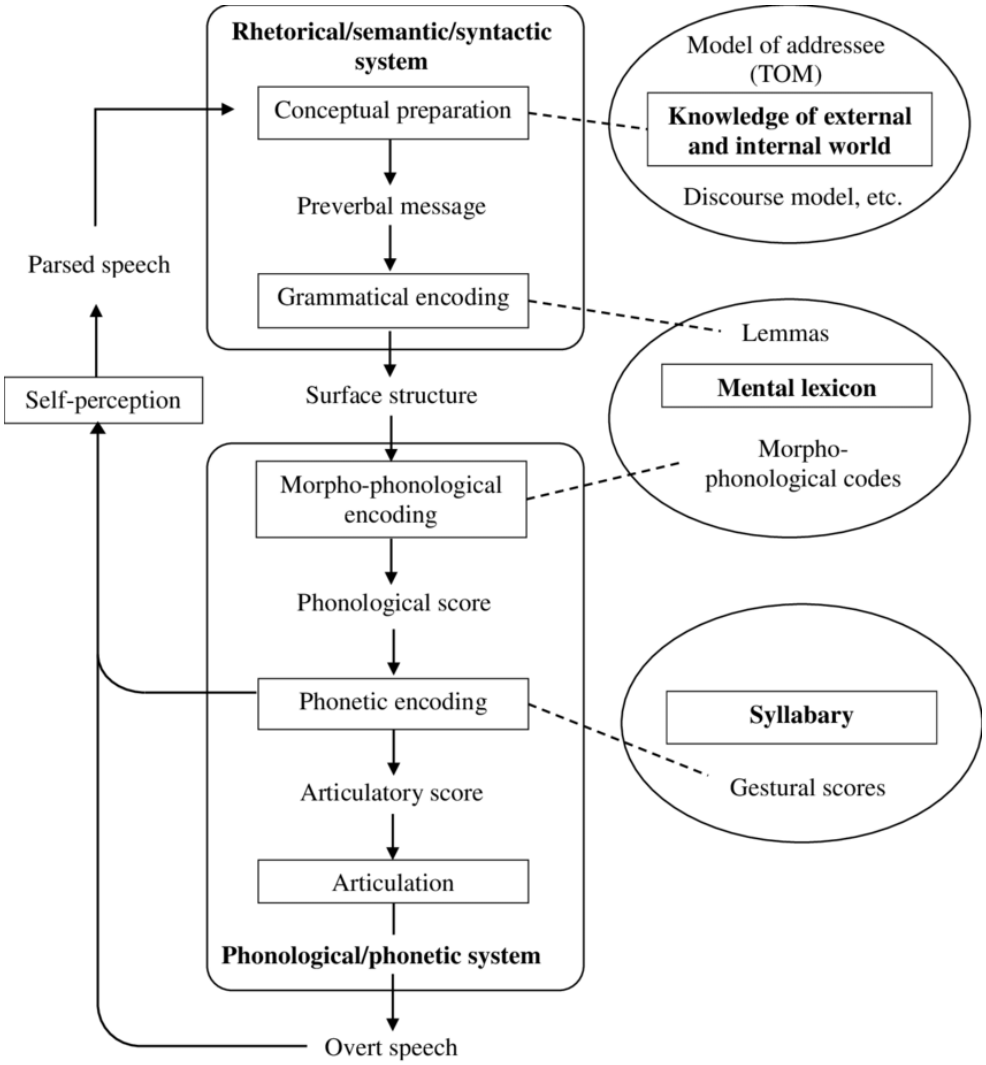


Figure 2.2 A blueprint of the speaker (Levelt, 1999: 68)

Model transformation after the Blueprint is exemplified as accommodating for the speech process of bilingual speakers. De Bot (1992) constructed a bilingual speech production framework based on Levelt (1999a), but reiterated the classification of code-switching and cross-linguistic inferences mentioned by Nortier (1989). Intended, situationally motivated, or contextual code-switching have helped justify the co-existence of different subsystems. With conceptualizer and macro-planning phase being non-language specific, speakers are in possession of only one lexicon for lexical element storage. However, the connection between lemma and form characteristics are not one-to-one for bilingual speakers, echoing with the assumption that different languages have their own formulators. Projecting the theories of L1 speech processing and production to L2 speech, Kormos (2006) depicted an integrated bilingual speech model that further modified Levelt's Blueprint Model. Kormos' bilingual speech model postulates a unique space for L2 declarative rules in long-term memory, suggesting that L2 speakers at a lower proficiency level have grammatical and morphological rules stored at an individual region.

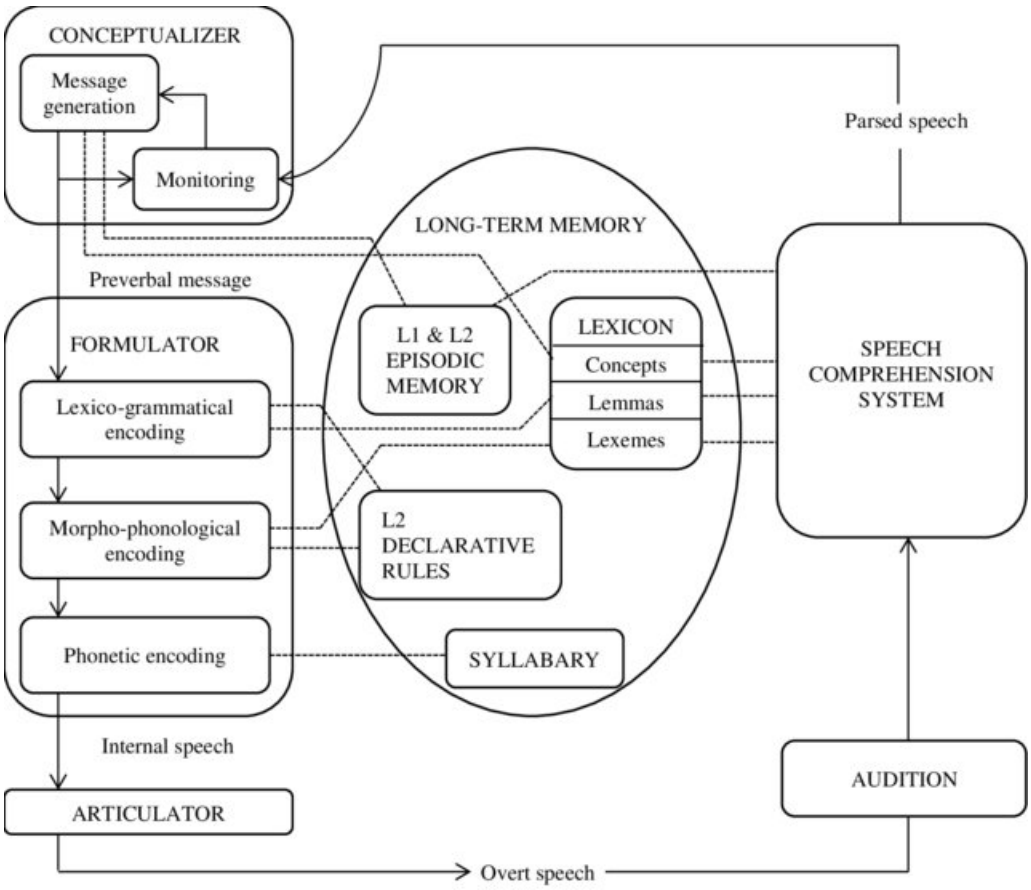


Figure 2.3 The Model of Bilingual Speech Production (Kormos, 2006: 168)

Cognitive models of speech, which emphasize the modular process of speech production and the function of memory, help categorize the main constructs in assessing speaking from the speakers' perspective. The conceptualizing and formulating stages involve the retrieval of vocabulary from various knowledge sources and mental lexicon. The final step, articulation, is often measured in terms of fluency and pronunciation. As has been described in Luoma (2004) as a "social" and "situation-based" activity, speaking is not easy to assess and involves stages in addition to those explained in cognitive models. Researchers have been paying attention to various criteria when evaluating and assessing L2 speaking, such as communicative competence (Canale & Swain, 1980; Hymes, 1972), interactional competence (He & Young, 1998; Kramsch,

1986; Young, 1999, 2008, 2011), or concretized components as language knowledge and strategic competence (Bachman & Palmer, 1996).

A wide range of specific features have been applied to evaluate a person's speaking proficiency, such as fluency, grammatical accuracy, and pronunciation. Understanding speaking, however, is complicated by the fact that performance and perception are combined in our roles as both speaker and listener, performer and audience. Given the normal acquisition of a first language, a speaker has mastered and thoroughly inhabits both roles. Speaking is mediated through the perceptions of the listener/rater (Yan & Ginther, 2017), and the larger part of speaking research has focused on speaker performance through listener perception and evaluation of speaking skills (e.g., segmental pronunciation, prosody, vocabulary, fluency, syntax, grammar). In the next sections of this literature review chapter, I will explain key constructs used to assess speaking, and the role of listener in L2 speech evaluation process.

2.2. Operationalization of Fluency as the Key Constructs Related with L2 Speaking

Fluency stands out as measurable constructs of speaking, as it directly mirrors the encoding process in speech production. Investigations of fluency have tended to focus on what Lennon (2000) characterized as fluency operationalized in the narrow sense, namely, corresponding to "the speed and smoothness of oral delivery" (p. 25) or the temporal variables associated with fluency that can be measured quantitatively. While the characteristics of temporal aspects of fluency have produced a rich and varied catalogue of research (Kormos & Rénes, 2004; Segalowitz & Freed, 2004; Segalowitz, French, & Guay, 2017; Towell, Hawkins, & Bazergui, 1996), the use of temporal representations alone to fully represent fluency has never been considered adequate. In contrast, Lennon's broad sense of fluency is associated with overall proficiency and corresponds to the commonsense notion of being a fluent speaker. With an effort

to distinguish individual variability in language performance, Fillmore (1979) discusses fluency as a dimension focusing on the oral language production end. Fluency is explained by four levels, which begins with an ability to talk with limited pauses and extends to higher requirements such as producing semantically condense sentences, having a command of speech appropriateness, and being capable of using a language creatively. Based on Fillmore (1979), who describes fluency as a phenomenon that covers pausing, coherence, appropriateness, and creativity in a broad sense, Ginther, Dimova, and Yang (2010) summarize the connections between L1 and L2 fluency:

> This broad sense of fluency extends into the domain of second language acquisition where the term is used to refer to mastery and ease of acquired second language performance (Faerch et al., 1984). First and second language domains are thought to converge when second language performance becomes 'nativelike' at high levels of proficiency (Chambers, 1997). (p. 381)

The narrow sense of fluency has been analyzed as a measurable construct manifested in language task performances. In a series of studies examining the characteristics of speaker fluency in relation to different tasks, Skehan (1996, 1998) investigates complexity, accuracy, fluency and lexical measures as overall representations of language performance. From a cognitive perspective, fluency indicates development of performance control with routinized and lexicalized language elements. As summarized in Skehan (2003):

> Regarding fluency, it is now increasingly accepted that finer grained analyses of fluency require separate measures of (a) silent (breakdown fluency), (b) reformulation, replacement, false starts, and repetition (repair fluency), (c) speech rate (e.g., words/syllables per minute), and (d) automisation, through measures of length of run (Koponen & Riggenbach, 2000). (p. 8)

Tavakoli and Skehan (2005) later combine sub-dimensions of fluency-related measures with speech rate and length of run, and named it speed fluency. Speech rate "refers to how fast

and dense the produced language is in terms of the time units". (p. 255). Length of run, as defined in Freed (2000), is continuous speech produced between pauses or hesitations (usually pauses of 250 milliseconds/0.25 seconds or greater).

The Blueprint system proposed by Levelt was further reinterpreted as a four-step procedure in Skehan, Foster, and Shum (2016), who argue speaking fluently relies on the assumptions of "knowing what you want to say" in the conceptualization stage, "having the means to say it" in the formulation stage, "not changing one's mind" during fluent delivery of message, and "anticipating problems effectively" for fluency maintenance (p. 97). A macro-level comparison was made between discourse dysfluency and clause dysfluency, determined by the location of breakdown pauses and repair strategies. Longer end-of-clause pauses are regarded as indicative of discourse dysfluency and is associated with the conceptualizer, while pauses occurring within a clause are associated with clause dysfluency and associated with the formulator and articulator.

Segalowitz's approach derives in part from his observation that "Despite several decades of work, researchers have not discovered universally applicable, objective measures of oral fluency" (Segalowitz 2010, p. 39). Indeed, fluency research has been characterized by considerable variation in operationalizations and measurement: speed (speech rate, articulation rate, phonation-time ratio, and mean length of run); pausing (the number, duration, and location of silent and filled pauses); and repair (false starts, repetitions, and reformulations).

Indices related with the measurement of utterance fluency, when reorganized with Tavakoli and Skehan's framework of speed fluency, breakdown fluency, and repair fluency, include examinations of filled pauses and unfilled pauses (Kormos, 2006; Kowal, O Connell & Sabin, 1975; Riggenbach, 1991), length of pauses (Goldman-Eisler, 1968; Kormos & Dénes,

2004; Towell et al., 1996), repeat, reformulation, false starts, and other disfluencies in speech (Freed, 1995; Hieke, 1985; Tonkyn, 2012), rate of speech (Blake, 2006; Freed, 2000), mean syllables per run (Ginther, Dimova, & Yang, 2010), and mean length of syllables (Bosker, Pinget, Quené, Sanders, & De Jong, 2013). More detailed explanation of utterance fluency measures and related research backgrounds is presented in Appendix E.

2.3. Measurement of Utterance Fluency

Utterance fluency is embodied by speakers' performance at the end of all the modular models presented, where the ultimate speech delivery is clearly measurable in terms of objective indices. In other words, measurement of utterance fluency characterizes the operational definition of fluency. Blake (2006) proposed three subcategories of fluency: pause-related variables, quantity-related variables, and repair-related variables. While pause-related variables involve identifying pauses and determining lengths of pauses, quantity-related variables calculate speakers' delivery speed. Repair-related variables demonstrate strategies used by speakers to cope with language output disruptions and disfluencies. The typology of fluency measures corroborates with the framework proposed by Tavakoli and Skehan (2005): breakdown fluency measured by pauses, speed fluency related with speed calculation, and repair fluency demonstrated by speaker-devised strategies.

Segalowitz (2001) extended understandings of fluency by explicating and differentiating cognitive fluency from performance fluency. He further categorizes performance fluency as utterance and perceived fluency, highlighting the roles of both speaker and listener. Referring to the operational mechanisms underlying performance fluency, cognitive fluency represents the ability to marshal resources needed for communicative purposes. Utterance fluency, interpreted

as an observable phenomenon closely related to Lennon's narrow sense of fluency, is measured through linguistic features related with quantity and speed of speech.

Segalowitz (2010) further clarified the definition of cognitive, utterance, and perceived fluencies: Cognitive fluency is described as the ability to process and make adjustments during the speech planning phase, assemble utterances and express individual interpretation with linguistic resources, along with the capability to realize sociolinguistic functions of a language. The features of utterance fluency concretize Lennon's earlier narrow sense of fluency. Utterance fluency features includes speech rate, hesitation and pausing phenomena, which are mostly categorized as speed and repair fluency in Tavakoli and Skehan (2005). Perceived fluency, however, stems from the listeners' perceptions.

Segalowitz (2016) explained the purpose of distinguishing utterance fluency from cognitive and perceived fluency and suggests reducing utterance fluency measures in research settings Studies of L2 fluency need to focus on performance features that reliably indicate speakers' ability to assemble speech. Segalowitz, French, and Guay (2017) defined utterance fluencyas "more narrowly in terms of temporal and hesitation phenomena that characterize the fluidity of speech delivery" (p. 92). While sifting through potential fluency markers, Segalowitz et al. (2017) followed three consecutive steps: identifying basic and quantified features that define utterance fluency, excluding measures that are mathematically transformed from existent indices, and eliminating those variables that either correlate too strongly or too weakly with others. A core set of utterance fluency measures was identified, which includes number of syllables between silent pauses, seconds of phonation (i.e. time of speaking) between silent pauses, pruned articulated syllable duration in milliseconds (i.e. 60, 000 ms divided by syllables per phonation minute), and mean silent pause duration. This core set of fluency variables helps

build specific and generalizable relationships between utterance fluency and cognitive fluency, as cognitive fluency is the underlying controlling system for speech production. A clearer association between L2 cognitive fluency and utterance fluency also downsizes possible confusion between L2 language abilities and abilities associated with other, more general cognitive tasks.

Focusing on the objective measurement of utterance fluency, this section of literature review explains related measures frequently used in fluency studies. The classification of fluency is further categorized as raw frequency representation and ratio-based (or normed) calculation. In addition, breakdown fluency and speed fluency overlap in their joint nature of being a subset of temporal fluency, which is "measurable by the rate of speaking, the length of fluent 'runs' between pauses of a standard length, and the frequency, length and placement of pauses"(Tonkyn, 2012, p. 225). A search in temporal variable used for breakdown fluency and speed fluency will fulfill the following research purposes:

(a) Identifying measures to be selected for creating profiles for L1 intermediate advanced and advanced L2 English speakers;

(b) Investigating additional dimensions of fluency, or more specifically, the influence of lexical factors on linguistic profile representation.

2.3.1. Quantitative Measures of Utterance Fluency

Utterance fluency is composed of breakdown fluency, speed fluency, and repair fluency. Early investigations of breakdown fluency, or periods of silence in speech production, were termed studies of pausology—"the behavioral investigation of temporal dimensions of human speech" (O'Connell & Kowal, 1980, p. 8). However, detailed descriptions of pauses or breakdown fluency focusing on filled pauses and unfilled pauses occurred even earlier, the latter

of which were defined by the duration of non-speech intervals in Maclay and Osgood (1959). Pauses were also categorized by Goldman-Eisler (1968) as: a) hesitation pauses unrelated to articulatory processes, or disfluency that interrupts spontaneous speech; b) grammatical pauses and non-grammatical pauses depending on the location they took place; or c) breathing pauses and non-breathing pauses in the phase of expiration for new air. Interestingly, Goldman-Eisler's investigations were undertaken to identify the characteristics of speech produced by clinically depressed subjects as an aid in diagnosis. A variety of pauses in linguistic domains have been identified as juncture pauses that occur at clause boundaries (Hawkins, 1971), rhetorical pauses (Deese, 1980) used to enhance rhetorical effects, or lexical and non-lexical pauses (Dörnyei & Kormos, 1998). Lexical pauses occur with fillers or gambits such as "well" and "you know", while non-lexical pauses are moments of silence or sound lengthening.

The length of silence or pauses has been a critical research question, especially when speed fluency and other ratio-based variables are calculated by number, amount, or proportion of pauses. Goldman-Eisler (1958) identified pause of hesitation as a generic category, with the cut-off point set at 0.25 seconds in a range extending to 6.0 seconds. Hieke, Kowal, and O'Connell (1983) held a discussion over the identification and location of pauses, as silent pause is an analytic unit in language production. Although a cut-off point between 0.2 and 0.3 is conventional in use when pauses are the object of research, more diverse measurement ranges starts from shorter than 0.1 second to longer than one second or even 2 seconds (Levin & Silverman, 1976; Siegman, 1979). More information can be found in Appendix E-1 (Breakdown Fluency Measures Related with Filled Pauses) and Appendix E-2 (Breakdown Fluency Measures Related with Unfilled Pauses).

In comparison to breakdown fluency, which is centered around the position and length of pauses, speed fluency reflects the quantity of speech production through its ratio against a certain time period. Common representations of speed fluency measurement include Mean Length of Runs (MLR), Speech Rate (SR), and Articulation Rate (AR), all of which focus on the number of syllables produced within a fixed time frame. One research purpose for speed fluency measurement is to explore the connection between utterance fluency and speakers' language proficiency level. Ginther, Dimova, and Yang (2010) reported the correlational results of 15 temporal fluency variables and the examinees' holistic oral proficiency level as measured by holistic scores on the OEPT, the same instrument investigated in this study. Mean Syllabus per Run ($r = 0.72$), Speech Rate ($r = 0.72$), and Articulation Rate ($r = 0.61$) were reported as having the strongest relationships with OEPT holistic scores.

Further investigation has shown dissimilar distinguishing functions among the three variables, all of which are associated with a steady growth of the examinees' language proficiency. Mean Syllables per Run (MSR) efficiently differentiates test takers of the highest proficiency with those of advanced intermediate proficiency. Speech Rate (SR) possesses limited discrimination power when speakers reach the threshold of 200 syllables per minute, but differs significantly between two adjacent groups of Chinese examinees with lower language proficiency levels. In order to distinguish among adjacent proficiency levels more clearly, researchers need to take other factors other than fluency into consideration, such as speakers' use of vocabulary.

More information can be found in Appendix E-3 (Speed Fluency Measures) about various indicators of speed fluency and the use of pauses for related calculations. The role speed fluency plays in performances of speakers at the same language proficiency level, however, is

open for discussion. Keeping long runs or maintaining a high speech speed may trigger a concomitant effect of additional and related variables, e.g., vocabulary usage. Speaking at a fast rate with limited lexical variety, however, might be insufficient to indicate advanced language proficiency. Also, the combination of fast rate and high vocabulary complexity may possibly impact listeners' perceptions of other constructs, such as accentedeness. In this dissertation study, I hypothesized that fast speech rate and diverse vocabulary may contribute to raters' stronger perception of accent.

The nature of variables related with repair fluency, which are extensively used in conversation analysis and the assessment of interactional competence, are more qualitative in nature. According to Young (2011), repair is "the ways in which participants respond to interactional trouble in a given practice." (p. 430). Variables related with repair fluency include strategies such as repetition, restart, and self-corrections. Riggenbach (1991) categorizes restarts as retraced starts and unretraced starts. Retraced starts include repetition, insertion, as well as recasts of the original speech parts. Unretraced starts refer to "reformulations in which the original utterance is rejected". (p. 427) and are equivalent to false starts (Rossiter, 2009). More detailed classification of repair-related variables also includes a variety of speech markers, such as cut-offs, prolonged sounds (lengthening vowels for instance), as well as incomplete words and phrases are described in Blake (2006). In most empirical studies, repair-related variables such as repetition and self-correction are transformed to ratios (De Jong, Groenhout, Schoonen, & Hulsijn, 2015; Huensch & Tracy-Ventura, 2017). Raw frequencies of repetitions and corrections are divided by time or a specific word count. Appendix E-4 (Repair Fluency Measures) enlists a collection of repair-fluency variables, which are frequently used in related research.

2.3.2. Fluency Variables Selected for Linguistic Profiles of Advanced Intermediate and Advanced L2 English Speakers

In this dissertation study, I decided to include three variables: Speech Rate (SR), Pause Rate (PR), and Mean Syllables per Run (MSR). Two of the four variables, Pause Rate (PR) and Mean Syllables per Run (MSR), were also included in the study of Segalowitz et al. (2017). Segalowitz et al. (2017) calculated number of syllables between silent pauses and mean silent pause duration, which correspond to Mean Syllables per Run (MSR) and Pause Rate (PR) in this study. I also select Speech Rate (SR) in this study as the variable is a representation of speed fluency.

Speech Rate (SR) and Pause Rate (PR) represent speed fluency and breakdown fluency respectively, the inclusion of which provides a well-rounded representation of utterance fluency as a measurable construct. Mean Syllables per Run (MSR) was also selected in this study. As a composite variable that integrates both breakdown fluency and speed fluency (Bosker et al., 2013; De Jong et al., 2015; Tavakoli, Campbell, & McCormack, 2016), Mean Syllables per Run (MSR) shows to be a key variable that strongly correlates with speakers' overall language proficiency or language development (Ginther, Dimova, & Yang, 2010; Segalowitz et al., 2017). The selection of temporal variables is expected to provide a comprehensive description of fluency as a construct, where both speech delivery speed and pauses are closely investigated as well.

Selecting from short-listed quantitative fluency features would also benefit cluster analysis as a statistical method. Cluster analysis explores data structure based on multiple quantified variables, and divides data points into groups with homogeneous numeric values. Involving a list of variables that are highly correlated with each other is counterproductive to efficiently extracting representative files. For example, speech samples having fast Speech Rate

(SR) may also have high values in Articulation Rate (AR). Including the two variables simultaneously will result in clustered groups that are dominantly high or low in both indices. In this situation, all the variables are subject to correlation examinations before cluster analysis is conducted, so that variable redundancy and collinearity will be avoided. Researchers need to search for a compromising middle ground while selecting variables that represent a measurable construct. The variables to be included need to embody different facets of a measurable construct. Meanwhile, statistical concerns also call for careful consideration.

The relationship between fluency and vocabulary is another key question to be addressed in this dissertation. Temporal fluency variables listed for gauging utterance fluency might constitute only partial explanation for smooth delivery. From a cognitive stance, automatic and controlled processing in cognitive fluency have been explored in tasks associated with word recognition (Segalowitz, 2001; Segalowitz, Watson, & Segalowitz, 1995). Word recognition progresses in a ballistic and stable fashion. Once the task is initiated, it draws little resource from other ongoing activities and cannot be stopped before completion. Faster performance thus involves a mix of automatic and controlled components simultaneously. Reduction in time consumed may be due to practice effect, which does not necessarily lead to speed increase in a blend of automatic and controlled components. In other words, a characteristic of an automatic process is not only that it is fast, but that it is also reliably fast. This argument is effective in selecting linguistic features in addition to narrowly defined utterance fluency. Speed alone does not possess enough power to validly indicate automatic or controlled processing and bears the following implications.

First, the measurement of fluency, or overall oral language proficiency, is not confined to examining variables directly related to speakers' delivery speed. Investigations from cognitive

perspectives have focused on disintegrating the composite structure of speaking proficiency, where linguistic knowledge and processing skills are both emphasized (De Jong et al., 2012a; Koizumi & In'nami, 2013). Knowledge of vocabulary has been consistently demonstrated to be a valid predictor of speaking proficiency. Broader vocabulary knowledge contributes considerably to speed fluency, as it empowers speakers at the formulator stage with faster speed gaining access to lexical resources. Second, in terms of representing fluency through temporal variables, a combination of speech rate and speech content is needed for a listener's perception of the speaker's level of fluency. Among the temporal variables describing performance fluency, ones that combine speech speed and quantity deserve more attention in measuring the construct (Ginther, Dimova, & Yang, 2010).

The social nature of speaking demands consideration of listener perception, which bolsters selection of fluency variables that have the greatest impact on listeners' evaluation. Based on the cognitive and componential view of L2 speaking proficiency, research has been conducted to disentangle the relationship between utterance fluency and perceived fluency. Segalowtiz (2010) introduced perceived fluency, which "has to do with the inferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency." (p. 48). The separation of perceived fluency from cognitive fluency and utterance fluency fully instantiates the role of listeners in communication. Utterance fluency can be manifested through objective measurement but is also inevitably connected with the listener's judgment. Bosker et al. (2013) investigated the contribution of pauses, speed, and repairs to perceived fluency of L2 Dutch speech. Untrained raters were asked to evaluate overall fluency of L2 speech, which is also measured by sets of acoustic fluency features. Fluency ratings were mostly explained by breakdown fluency and speed fluency measures, followed by repair fluency. Different groups of

raters were also asked for subjective evaluation of pauses, speed of delivery, and the use of hesitations and corrections. When subjective ratings on specific fluency aspects were used to predict overall fluency, ratings for pausing explained most of the variance, closely followed by those for speed fluency. However, the major contributing role of speed fluency and breakdown fluency is also attributed to the strong correlation between the two, which aligns with the arguments presented in Derwing, Rossiter, Munro, and Thomson (2004) and Rossiter (2009). Again, variables related to delivery speed and pausing need to be prioritized when selecting variables that represent fluency, and these are subject to examination of correlation prior to statistical analyses.

A few arguments can be made after synthesizing the research that bridges utterance fluency and perceived fluency. Discarding variables that are at the core to fluency analysis may lead to inadequate representation of the construct, which justifies the inclusion of breakdown fluency and speed fluency variables (i.e. Speech Rate (SR) and Pause Rate (PR)). Also, variables included in the study need to be functional in differentiating between proficiency levels, especially speakers rated on the higher end of the scale. As a composite variable that weighs in both speed fluency and breakdown fluency, Mean Syllables per Run (MSR) will be used in this study to evaluate advanced L2 English speakers' performance, together with Speech Rate (SR) and Pause Rate (PR). Also, with reliably fast speech delivery becoming a connotation of fluency, the measurement of speech content is tightly connected with another important construct—vocabulary.

2.4. Vocabulary—a Dimension in Interaction with Fluency

In speaking assessment, the relationship between vocabulary and fluency has strong and convincing precedents, and the contribution of vocabulary knowledge to cognitive fluency has

ensured its critical place. Pawley and Syder's (1983) now classic article discussed the puzzle of "nativelike selection" and "nativelike fluency" and argued that lexis, especially lexicalized sentence stems or collocations, play an important role in accounting for nativelike fluency. They questioned the assumption that generative rules alone could account for L1 linguistic competence, especially with regards to fluency, and argued for the critical place of nativelike selection (fixed, highly frequent phrases or collocations) in support of nativelike fluency. The production of long stretches of language output require not only high levels of automaticity but also selection of expressions that sound natural and idiomatic. They argued that nativelike selection is required to facilitate both speakers' and listeners' real time production and processing and assists in explaining the puzzle of nativelike fluency.

The positioning of vocabulary assessment has been discussed in Bachman and Palmer's (1996) explanatory table of language knowledge, where vocabulary is listed as a subcategory of grammatical knowledge. Read (2000) argued that it is a narrow view to consider vocabulary "as a stock of meaningful word forms that fit into slots in sentence frames" (p.5), and an obvious place for vocabulary is the category of sociolinguistic knowledge. Integrating vocabulary into the discussion of fluency, however, is necessitated by the construct of speech production, where the lexico-grammatical encoding precedes overt speech production (Levelt, 1999a; Kormos, 2006). Hilton (2008) points out that "few studies address the contribution of lexical knowledge to spoken fluency" (p.153). Not only is vocabulary knowledge positively correlated with temporal fluency measures such as words per minute, investigation of disfluencies also reveals that most hesitations are followed by a lexical error or an overtly marked lexical search. The impact of lexical knowledge or lexical competence on fluency is thus nonnegligible.

In addition, the definition of fluency as a measurable construct is inseparable from the sources from which it forms and develops. Upon stratifying the four levels of fluency, Fillmore (1979) articulates that the source of fluency is "speaker's knowledge of fixed linguistic forms". Connection between lexical usage and other dimensions of oral language production in the complexity, accuracy, and fluency (CAF) framework has also been examined. In Tavakoli and Foster (2011), lexical diversity was used as a separate domain in evaluating test performances in addition to complexity and fluency. The interpretation of lexical diversity, or lexical proficiency development in a broader scope, remains questionable whether being housed independently in the realm of complexity, accuracy, and fluency, or as an integral part of the three major dimensions. Besides, arguments have been made that fluency should not be confined within phonological dimensions only, as deleting disfluency markers from responses of lower-level proficiency speakers only brings limited gain in the final speech evaluation score (Cao, 2014). Lexical use is a highly distinguishable element displayed by various linguistic profiles, and the influence of vocabulary becomes prominent when speaking tasks are set in academic contexts. For instance, disparities in using academic vocabulary is highly possible to result in perception differences of fluency, accentedness, or even overall language proficiency. This section of literature review covers indices that measure vocabulary performance and justifies the variables to be used in establishing linguistic profiles.

2.4.1. Terminologies used in Vocabulary Studies

Measuring the productive use of vocabulary is connected with approaches in understanding a series of terminologies: lexical diversity, lexical variation, lexical sophistication, lexical density, and lexical richness. Malvern and Richards (2002) describes lexical diversity as a component of lexical richness, reflecting "the variety of active vocabulary deployed by a speaker

or writer" (p. 87). Daller, van Hout, and Treffers-Daller (2003) interpreted lexical richness as a construct connected with vocabulary size, which is best known to be measured by Type Token Ratio (TTR). TTR is an index widely used in vocabulary assessment, where the number of different lexical items (types) is divided by the total number of words (token). In the classification system of Read (2000), however, TTR is applied to evaluate lexical variation. Both lexical variation and lexical diversity have word repetition rate as a core measurable construct, and higher proficiency language learners are expected to have a lower repetition rate in their vocabulary usage. In comparison, lexical sophistication is related with word frequency. The use of low-frequency vocabulary is an indication of lexical sophistication. Lexical density is embodied by the percentage of content words compared to grammatical/function words.

Information and explanation of all the terms can be found in the work of Jarvis (2012), who calls for theoretical motivation and a well-developed model in understanding vocabulary knowledge. Jarvis (2013b) provides information about the historical development of the notion of lexical diversity, which can be deconstructed to six properties such as variability, disparity, and evenness. The variety and range of words used by language learners, for example, would have varying effects on listeners' judgment on the construct. Again, often measured through the relationship between type and token, lexical diversity has word repetition at its foundation. The term of lexical diversity is used interchangeably with lexical variation in studies such as Engber (1995). Lexical richness, which has also been interpreted as a synonym of lexical diversity (Daller, van Hout, & Treffers-Daller, 2003), is more often used as a hypernym that includes lexical sophistication and other constructs related with vocabulary usage (Read, 2000).

Different understanding of terminology and varied operationalization mechanisms still exist in measuring vocabulary usage. Subcategories of lexical richness, however, need to be

clearly defined in empirical studies, especially when quantitative approaches are adopted to assess vocabulary knowledge. To restate the status of vocabulary usage in establishing linguistic profiles of high proficiency L2 English speakers, I will include the following questions in this section of literature review:

(a) What role do proficiency development models play in the assessment of vocabulary?

(b) What are the representative subcategories in predicting language learner's lexical competence in speaking performance?

(c) What are the most commonly used measures to assess subcategories of vocabulary usage?

2.4.2. Lexical Proficiency Development Models

Selecting quantitative indices of measuring vocabulary usage is closely connected with the trajectory of vocabulary development. In other words, researchers would always want to investigate the constructs that efficiently capture learners' progress in lexical competence development. Richards (1976) provided a series of assumptions in measuring lexical competence, where knowing a word suggests understanding its morphological, syntactic, semantic, as well as socio-cultural association. In the context of measuring reading comprehension, Anderson and Peabody (1981) deconstructs knowledge of word meanings into breadth and depth. Breadth refers to the quantity of words a person knows, and depth stands for the degree to which a person understands the word. Depth of vocabulary knowledge is further defined by Read (1993) as "the extent to which learners were familiar with the meaning and uses of a target word." (p. 43). Upon dividing vocabulary knowledge into receptive and productive, Nation (1990) lists three aspects of knowing a word: knowing the form of word such as spelling and pronunciation, knowing the meaning of a word such as its referents and associations, and knowing the function of word such as its collocations and use constraints.

Definitions of global characteristics for lexical competence can be found in Meara (1996a) as vocabulary size and vocabulary organization skills. While vocabulary size is measured by related testing tools, vocabulary organization is a synonym of semantic association network and the connectivity among different words. Meara's (1996a) description of lexical competence is related to Read's (1993) analysis of a word association test that assesses vocabulary depth. Three types of relationship exist between the stimulus word and its associates: (a) paradigmatic that emphasizes on semantic synonym; (b) syntagmatic that focuses on collocation, and (c) analytic that accounts for the possibility of the associate being part of the stimulus word's dictionary definition. In comparison, the three dimensions of lexical competence proposed by Henriken (1999) are presented as a continuum to describe lexical proficiency improvement: (a) partial to precise knowledge, (b) depth of knowledge, and (c) receptive to productive use ability.

Corresponding to the growth of vocabulary knowledge, vocabulary assessment techniques have been used in different contexts to measure lexical proficiency growth, Read (2000) proposes three aspects for vocabulary assessment: (a) discrete and embedded, where discrete test takes vocabulary as a distinct construct separated from other components of knowledge, and embedded assessment integrates vocabulary knowledge into other language constructs; (b) selective and comprehensive, where selective means measuring specific vocabulary items and comprehensive measures are inclusive to the entire vocabulary content; and (c) context-independent and context dependent, where contextual information is deprived or provided for test takers' to produce expected responses.

With lexical competence expanding and shifting from receptive vocabulary to productive vocabulary, lexical assessment is focusing on learners' actual use of vocabulary in their written

and oral speech output (Melka, 1997; Nation, 2001). Read (2004) defines three lines of development in measuring vocabulary depth of L2 acquisition: precision of meaning, or commanding specific and elaborated knowledge of words' meaning; comprehensive knowledge of a word, or understanding the meaning of words in syntax, morphology, collocation, and pragmatics; network knowledge, or the ability to locate words in a lexical network and differentiate them from other related words. More detailed definition about depth is revealing the overlapping area between depth and breadth. As Read (2004: 218) explains: "…if learners demonstrate 'more advanced' kinds of knowledge of particular words, we can assume that they have acquired 'basic' knowledge of those same words."

Meara (2005) describes lexical competence through "vocabulary size, depth of vocabulary knowledge, and the accessibility of core lexical items." (p. 271), which suggests that breadth and depth are interpreted in combination with other dimensions in terms of explaining the enhancement of lexical competence and growth of lexical proficiency. Along with the addition of fluency, Daller, Milton, and Treffers-Daller (2007) proposes a three-dimensional space to evaluate language learners' vocabulary knowledge. The amount of words a learner knows is measured by breadth. Depth is defined by learner's confound understanding of a specific word. Fluency is connected with the level of automaticity when a learner uses the word. The model is a recognition of possible intersections among the three dimensions.

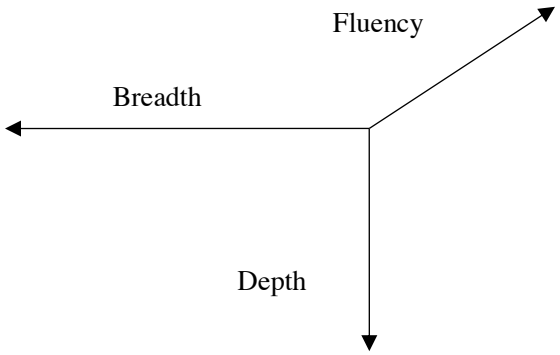


Figure 2.4 The lexical space: dimensions of word knowledge and ability (Daller, Milton & Treffers-Daller, 2007: 8)

Crossley, Salsbury, McNamara, and Jarvis (2011a) further clarified the definition of lexical proficiency, which nods to the lexical space model proposed by Daller et al. (2007) and consolidates the status of fluency in the model:

> Generally speaking, lexical proficiency comprises breadth of knowledge features (i.e. how many leaners a learner knows), depth of knowledge features (how well a learner knows a word), and access to core lexical items (i.e. how quickly words can be retrieved and processed; Meara, 2005) (p. 182).

A variety of natural language processing tools have been applied to analyze vocabulary breadth and depth. By connecting computational indices with human ratings of lexical usage, studies with a computational linguistics approach have demonstrated language characteristics of advanced L2 learners' performance. L2 language users with higher linguistic proficiency have showcased higher lexical diversity, a wider range of hypernymy levels, and more frequent use of abstract words (Crossley, Salsbury, & McNamara, 2009, 2012; Crossley, Salsbury, McNamara, & Jarvis, 2011b). Other predictive features of lexical proficiency include word frequency, word association, and word familiarity (Crossley & Salsbury, 2010).

Fine-tuning of the lexical proficiency model foregrounds dimensions that evaluate learners' lexical competence through surface linguistic features. Crossley et al. (2012) explained

the significance of both lexical diversity measures and word frequency information, as they efficiently assess language users' lexical competence by estimating the breadth of their vocabulary knowledge. You (2014) proposed the Cube as a new three-dimensional model for lexical proficiency measurement, synthesizing and combining theories of Henrikson (1999) and Daller, Milton, and Treffers-Daller (2007) about lexical proficiency development. Starting from the point where a learner's lexical space grows, the Cube has three axes demonstrating breadth, depth and fluency. The model establishes sides between every two axes, representing evolving dimensions in vocabulary proficiency development. Variety in production is located between the two axes of breadth and fluency. Sophistication in lexical production grows out of fluency and depth, while reception is a dimension established by depth and breadth.

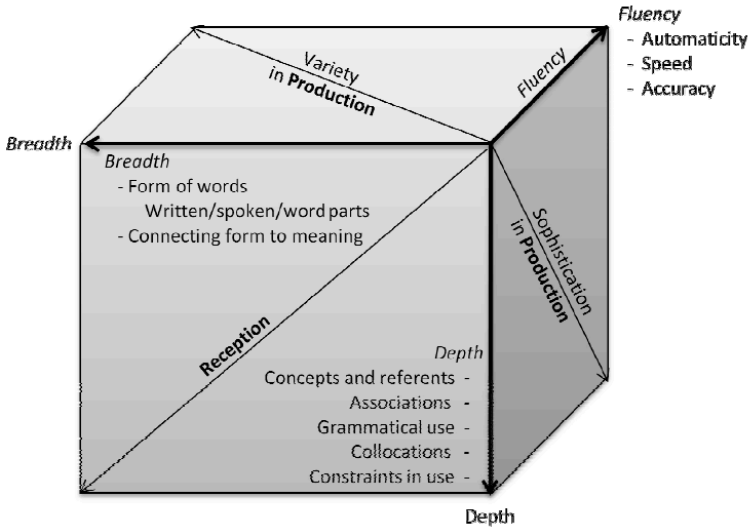


Figure 2.5 The Cube (from You, 2014: 15)

Emphasizing interactions among the three dimensions in language production tasks, the Cube represents a combined measurement of both lexical diversity and the Lexical Frequency Profile (LFP) in L2 oral English. In spite of being a critical index in assessing vocabulary knowledge, Lexical diversity is insufficient to offer a well-represented picture of lexical usage,

or vocabulary breadth in particular. While lexical diversity calculates the relationship between word type and token, the lexical frequency profile (LFP) emphasizes word frequency. An example as "The bishop observed the actress" was used by Schmitt (2010) to further explicate the relationship between the two. Sharing the same type-token analysis results with "The man saw the woman", the first sentence exhibits a completely different degree of sophistication. The expansion of lexical proficiency development models has confirmed that lexical richness is multifaceted. Lexical diversity remains an important index in assessing language learners' vocabulary knowledge as a surface linguistic feature, but also leaves gaps for other approaches that represent lexical proficiency growth. Measurement indices applied for evaluating vocabulary breadth will be further explored in this dissertation. In Appendix F, previous studies integrating quantified measures have been categorized and examined. More connection will be established between lexical factor and other linguistic profile information of L2 English speakers, along with possible variations caused by different experimental contexts.

2.4.3. References of Vocabulary Breadth Measures

The discussion of vocabulary breadth measures cannot be separated easily from the concept of lexical richness, which is explained as a general term inclusive of diverse measures of productive vocabulary. Being a subconstruct in describing the breadth of vocabulary knowledge, lexical richness greatly weighs in assessing language learners' lexical competence, or even overall language proficiency. Various indices have been applied to quantify lexical richness for measurement purposes. Reasons in support of using specific lexical richness indices, however, are usually implicit and not completely explicated. To understand the functioning mechanisms of quantitative lexical measures and their empirical application, I include two questions in this section of literature review:

(a) How is lexical richness quantified in the measurement of vocabulary knowledge?

(b) How are the quantitative lexical measures used in specific research contexts?

Lexical richness, as encompassing as it is in the lexical proficiency models, are disintegrated into quite a few aspects when applied in vocabulary measurement. Previous research is categorized in Appendix F, where measurement constructs are defined with specific mechanisms of calculation and evaluation of functioning efficiency. In general, lexical richness covers a) lexical diversity, or the variation of words; b) lexical sophistication, or the use of words at an advanced level; c) lexical density, or the ration between content words and function words.

Among the various empirical studies surveyed, I coded the quantification indices in Appendix F with the following four measured subconstructs lexical richness, lexical density, lexical diversity, and lexical density. Researchers usually measure multiple subconstructs in their studies simultaneously, which results in the appearance of several indices at the same time. I also catalogued the references based on research contexts. Measurement of oral and written vocabulary proficiency development takes place in both L1 and L2 settings. A more detailed coding framework is exemplified by L1 K-12/Speaking, L1 K-12/Writing, L2 K-12/Speaking, L2 K-12/Writing, L2 Adult/Speaking, L2 Adult/Writing. The last two categories, L2 Adult/Speaking and L2 Adult/Writing, include studies conducted in higher education institutes where adult second language learners were participants.

Annotations for each study also include explanation for the calculation mechanism of the indices, along with researchers' commentaries on their efficiency. The purpose of creating an inventory of lexical measures is two-fold. First, researchers can identify the essential facets to cover when establishing a profile of language learners' use of vocabulary. Second, detailed explanation and solid justification can be provided when quantitative measures will be used.

A number of themes are identified from the research that made use of lexical measures, which can be concluded as summarized answers for the research questions of this section. First, lexical richness tends to be a more general term covering all other related constructs. Among the quantitative measures of lexical richness, lexical diversity remains a critical aspect and includes ratio-based variables. As a genuine form of ratio between type and token, Type Token Ratio (TTR) is the foundation from which a series of other lexical diversity measures are derived, such as Malvern and Richard's $\mathcal{D}$. In addition, lexical diversity is interpreted and operationalized from a micro level, where researchers have articulated specific types of words to be measured (e.g. rare words, functional words). This developmental trend is exemplified by the calculation of Rare Word Diversity (Malvern & Richards, 2009) and lexical diversity of separate functional word categories (Treffers-Daller, 2009). Thirdly, the procedures for lexical diversity calculation are gearing towards probability-based approaches from ratio-based explanations. Lexical diversity measures derived from TTR will be discussed in the next section, where curve-fitting and probability-driven methods in measuring lexical proficiency are explained.

2.4.4. Quantitative Measures of Lexical Diversity—a Brief Review

In research related with L2 writing, lexical diversity represents a similar construct as lexical variation and lexical variety (Engber, 1995). Considered as a type of measure that reflects learners' breadth of vocabulary knowledge and language proficiency, lexical diversity is often described through a relationship between token and type. Within the terminological system of lexical measurement, token refers to "the total number of words in a text or corpus" while type means "the number of different words" (Daller, Milton, & Treffers-Daller, 2007).

Calculations of lexical diversity are based on the relationship between type (class of words) and token (number of words). Type Token Ratio (TTR), which was proposed by Johnson

(1939, 1944) as an attempt to address sample-size dependency problem, is "calculated from a standard number of tokens from each text (e.g. the first 200 words)" (Jarvis, 2013a, p.91). An alternative is referred to as Mean Segmental Type Token Ratio (MSTTR) (Richards & Malvern, 1997), which is the average of TTR from multiple based on equally sized subsamples of a text.

Rather than solely focusing on the relationship between type and token, Vemeer (2000) recommends that the difficulty of words should be considered in later stages of language acquisition, preferably when the vocabulary size reaches 3,000 and above. The functioning of TTR, as explained in Daller, Van Hout, and Treffers-Daller (2003), is not capable of distinguishing between word types such as basic vocabulary and advanced vocabulary. Holding the principle that not all words carry equal weight, Daller et al. (2003) advocate measures that include a qualitative dimension that gives more insight into lexical aspects of language proficiency instead of direct quantification.

Another conspicuous reason for TTR's unsatisfactory or unconvincing performance is attributed to its sensitivity to text length. TTR is a valid measure for gauging vocabulary development in children as their vocabulary increases along their ability to produce texts. However, the ration reaches a ceiling once the ability to produce text reached a particular number of words, resulting in TTR's failure to discriminate. For more advanced adult second language speakers, the chance for a new word to appear drops lower as the text length increases, which results in its instability in measuring lexical richness (Daller et al., 2003). Problems incurred with text length are not fully addressed despite series of indices based on TTR adjustment, such as Carroll (1964)'s corrected TTR, Guiraud's R (1954, 1960), and Herdan (1960).

As a mathematical model describing Type Token Ratio (TTR), D-measurement, or Malvern and Richard's $\mathcal{D}$ introduced in Malvern et al. (1997, 2002) sets up a plot of TTR (y-

axis) against token N (x-axis). It measures lexical diversity by "matching the graph derived from a real language sample to the ideal curves of this model". Based on the work of Sichel (1971, 1975) in search of a formulation of the ideal curve, Malvern et al. (2004) developed a mathematical expression that applies for a small sample approximation: $\mathrm{TTR} = \frac{\mathcal{D}}{\mathrm{N}}[\left(1 + 2\frac{\mathrm{N}}{\mathcal{D}}\right)^{\frac{1}{2}} - 1]$. A larger $\mathcal{D}$ coefficient is accompanied by a higher curve, which signifies greater lexical diversity.

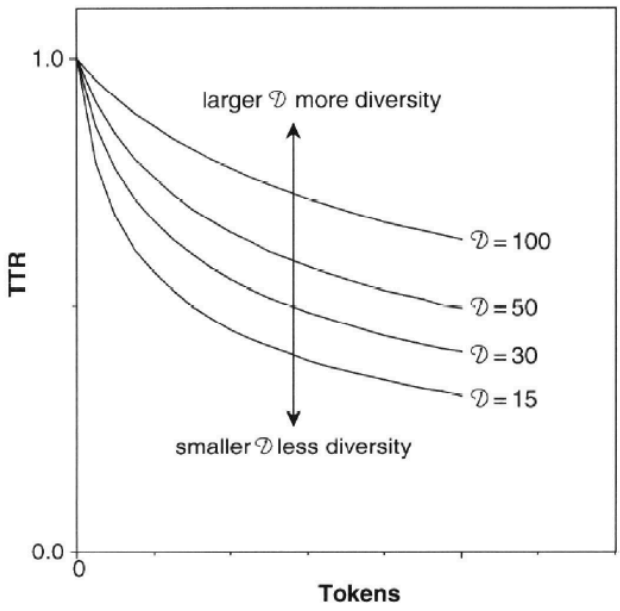


Figure 2.6 Ideal TTR versus Token Curves (from Malvern et al., 2004: 52)

Mirroring Simpson (1949)'s production of a precise probability of randomly choosing two individuals in the same category (the same word type) twice in succession, the coefficient of $\mathcal{D}$ provides the best curve after random sampling without replacement. As McCarthy and Jarvis (2010) explained, the procedure starts with taking 100 random samples of 35 tokens and calculating a mean TTR. Followed by repeating the procedure for 36 tokens and all up to 50, Malvern and Richard's $\mathcal{D}$ helps plot a random sampling TTR curve for the text. The theoretical

curve is produced by using the formula in Malvern et al. (2004), $\text{TTR} = \frac{\mathcal{D}}{\text{N}}[\left(1 + 2\frac{\text{N}}{\mathcal{D}}\right)^{\frac{1}{2}}\text{-}1]$, providing the best fit between theoretical curve and the randomly-sampling TTR.

In the calculation procedure for Malvern and Richard's $\mathcal{D}$, N stands for the number of tokens and the estimate for constancy. Malvern et al. (2004:59) point out that "$\mathcal{D}$ is a particular value for best-fit between the ideal curves and those derived from real transcripts over the standard range of points of the TTR versus N curve drawn by a standardized procedure." D-Measurement, which was substantiated to address the problem of sample-size dependency and thus presenting a robust index, is doubted by Jarvis (2002) and McCarthy and Jarvis (2007). Jarvis (2002) commented on the positivity of the curve-fitting approach and averaging TTRs of randomized tokens. However, by comparing Malvern and Richard's $\mathcal{D}$ (i.e. D-measurement) with four other lexical diversity measures (Herdan's C, Guiraud's R, Uber's U, and Zipf's Z) in modelling TTR curves, Jarvis (2002) suggested that Uber's U may be a more suitable lexical diversity measure for content-word texts. It remained questionable whether the lexical diversity of content words results in higher correlations with language users' written output. McCarthy and Jarvis (2007) argued that the Malvern and Richard's $\mathcal{D}$ value is highly correlated with the average TTR for all possible combinations of certain words, with the sum of probabilities (SOP) of each word being the essence for measurement. The coefficient $\mathcal{D}$ ended up converting lexical diversity to a new scale—changing probability to TTR and then generating a $\mathcal{D}$ coefficient value. However, an overcompensation of TTR with the text length occurred in the end, which leaves text-length dependency again at the core of problem.

McCarthy and Jarvis (2010) looked into the Measure of Textual Lexical Diversity (MTLD) in a following validation study. It is calculated "as the mean length of sequential word string in a text that maintains a given TTR value (0.72)" (p. 384). Each word is evaluated

sequentially for its TTR first. Factor count is conducted as the second step, which increases by 1 if a word has met the cutoff TTR value of 0.72. A partial factor count is also provided for the remainder of a lexical item, which is calculated as the range covered between 1.00 to 0.72. The final step in calculation of MTLD is completed when the total number of words divided by the total factor count.

The use of MTLD is supported with arguments that no remaining data would be discarded, meanwhile fully substantiating the concept of theme saturation. The gist is calculating the number of words it takes to reach an area, which is located prior to a point of stabilization. It is a point where neither repetition nor new type strings would affect the TTR trajectory. Validation results also show that there is no correlation between MTLD and text length, which helps contributes for arguments for its use. A brief history of index development would benefit research of the next phase: comparing the functional efficiency of different measures and include MTLD as the vocabulary measure for cluster analysis.

2.4.5. Quantitative Measures of Lexical Frequency Profiles

Lexical proficiency models such as the Cube (Daller et al., 2007; Henriksen, 1999; You, 2014) rely on a connection between lexical diversity and more qualitative investigation of word types. Nation (2001) mentioned that vocabulary learning should be directed to more specialized areas when learners' vocabulary size has reached a threshold. The discussion leads to pinpointing the importance of academic vocabulary instruction for the word type's high appearance frequency in academic English. As an argument against over reliance on Type Token Ratio (TTR), Vermeer (2004) suggests that more consideration should be given to the difficulty of a word, as more difficult words will not be mastered by speakers until advanced stages of language acquisition.

In addition to lexical diversity, Lexical Frequency Profile (Laufer, 1994, 1995; Laufer & Nation, 1995) is used to discriminate between learners of different proficiency levels. Lexical Frequency Profile (LFP) evolves from terms such as lexical originality, lexical density, lexical sophistication, lexical variation (often measured by indices such as Type Token Ratio), or lexical quality. Having been applied to less proficient language users as well as more advanced students, Lexical Frequency Profile (LFP) reference includes: a) the first 1,000 most frequently used words; b) the second 1,000 most frequently used words; c) academic vocabulary and d) the use of less frequent words.

The position of academic words in LFP research has been a focus for language instruction, especially English for academic purposes. The development of Academic Word List (AWL), or word lists identified as important when teaching English for academic purposes, is explained in Coxhead (2000) at length. Outside of the first 2,000 most frequently used English words, AWL words occur across a wide range of themes and more than 100 times in academic corpora. From a pedagogical perspective, Laufer and Nation (1999) stated that the distinction between high frequency and low frequency words can help instructors identify learners' language development stages. LFP is thus used in a controlled productive test to measure examinees' vocabulary growth.

Attention towards AWL in L2 English speakers' performance is also substantiated by You (2014). The study found that L1 Hindi speakers produced more words from the Academic Word List (AWL), which is even a more common phenomenon for test takers scored at the intermediate advanced level (OEPT 50) than advanced level (OEPT 60). The correlation between AWL frequency and OEPT final score reached 0.39, which is moderately strong. Meanwhile, Mandarin speakers across all OEPT score groups used a large number of words (tokens), more

different number of words (types), and made more frequent use of lexical items in the vocabulary list of the first 1,000 frequent words. In this dissertation, AWL remains to be the research focus. It is an assumption that speakers of higher-level proficiency will have a better command of more complicated vocabulary, especially those words associated with an academic context. However, counter argument has also been made that high-level English learners are more familiar with mare frequently used formulaic expressions.

Features of both utterance fluency and productive vocabulary will help establish linguistic profiles of English second language speaker in this study. Another critical aspect in assessing speaking, however, is connected with the ultimate step of the cognitive models of speaking—articulation. Chunks of language unit are assembled and syllabified, leading to an interactive phase that involves listeners' evaluation. The next section of literature review focuses on the measurement of accentedness, a construct inseparable from listener perception and social implications.

2.5. Accentedness Measurement

According to the definition provided by *A Dictionary of Linguistics and Phonetics*, accent is "the cumulative auditory effect of those features of pronunciation which identify where a person is from, regionally or geographically" (Crystal, 2008, p. 3). The identification of accent is largely based on segmentals, metrics (i.e. beats of a line of poetry) and pitch. As a salient aspect in communicative fluency, accent is linguistically defined as "encompassing phonetic and phonemic aspects of speech alongside intonation, pitch, rhythm, length, juncture, and stress" (Crystal, 2008, p.3). Linguistic features describing a person's speech accent often lead to sounding different, or as Derwing and Munro (2015: 5) clarify, accent is "a particular pattern of pronunciation that is perceived to distinguish members of different speech communities". The

perception and social functions of accent have resulted in a dynamic construct, with specific segmental parameters, suprasegmental demonstration, and socio-cultural implications all being contributing factors (Moyer & Levis, 2014).

2.5.1. Social Implications of Accentedness

Given both the historical precedents and the strength of the relationship between accent and identity, this section of literature review begins with a discussion of the social implications of accentedness. Being a fluent speaker of at least one language is so familiar a part of our identities that listeners reliably recognize an accent different from their own after listening to as little as 30 milliseconds of recorded speech (Flege, 1984). Listeners are also able to reliably identify L1 and L2 speakers of languages that they do not speak (Major, 2007), and the presence of an accent has been found to affect language processing strategies of children as young as 16 months of age (Weatherhead & White, 2018). Scovel (1988) remarks: "accent features are exceptionally salient, and as a result we're very good at detecting perceived outsiders on the basis of their speech patterns". (p. 477).

Speakers' identity, social status, and even personal traits are strongly influenced by how they sound to others. Research in social psychology has found that listeners attribute a variety of characteristics to speakers based on accent, including nationality, regional membership, ethnicity, and social class (Labov, 2006), intelligence (Lambert, Hodgson, Gardner, & Fillenbaum, 1960), social desirability (Kinzler & DeJesus, 2013), and suitability for employment (Kalin & Rayko, 1978). Being an L2 speaker increases the potential effect of negative attributions as perceptions of a foreign accent have been found to influence listeners' beliefs about L2 speakers' general communication skills (Hosoda, Stone-Romero, & Walter, 2007) and overall competence (Nelson, Signorella, & Botti, 2016). Furthermore, speakers identified as

having foreign accents have also been assumed less credible (Bourdieu & Thompson, 1991; Lev-Ari & Keysar, 2010; Livingston, Schilpzand, & Erez, 2017), less educated (Fraser & Kelly, 2012), and less intelligent (Anderson et al., 2007; Fuertes, Potere, & Ramirez, 2002). Given the association of identity and accent along with the potential bias of our accentedness-based attributions, it is only natural that applied linguists, language testers, and language teachers have paid a great deal of attention to accentedness.

Investigation of accentedness in this dissertation study is also associated with situational implications. Linguistic features demonstrated by higher level L2 speakers' performance vary from characteristics demonstrated by lower proficiency users. All of the L2 English speakers in this study scored 50 or above in OEPT and successfully passed the exam. Listeners/raters have reached an agreement that only minimal listener effort is required to understand their speech. Compared with other constructs of pronunciation, such as intelligibility, accentedness may be the most noticeable linguistic feature. In addition, the language development trajectory of higher level L2 English speakers is no longer monitored by classroom instruction due to their sufficient language skills. Bodies of literature, however, still reveal undergraduate students' negative reaction to international graduate teaching assistants' accent despite of their preparedness in linguistic knowledge (Kang, 2008; Kang & Rubin, 2014; Rubin, 2012; Rubin & Smith, 1990; Smith, Strom, & Muthswamy, 2005). As a phenomenon bearing rich social implications, accentedness can be better explained when its manifestation is investigated among high proficiency L2 speakers. In addition, related information generated from the linguistic profiles can help both speakers and listeners make adjustments accordingly.

2.5.2. A Trinity of Intelligibility, Comprehensibility, and Accentedness

The discussion of accentedness is deeply embedded with the concepts of intelligibility and comprehensibility. Few researchers have contributed more to the development of the concepts of accentedness, comprehensibility, and intelligibility than Tracey M. Derwing and Murray J. Munro (Munro & Derwing, 1995a, 1995b, 2011; Derwing & Munro, 1997, 2005, 2009). While accentedness, comprehensibility, and intelligibility were concepts long present in the literature (Abercrombie, 1949; Morley, 1994; Pennington & Richards, 1986), Derwing and Munro developed these themes in a series of related studies that left no stone unturned in their attempts to clarify the relationships involved. Discussions of accentedness now go hand in hand with the concepts of comprehensibility and intelligibility. In fact, the association between accentedness, comprehensibility, and intelligibility has so permeated the discussion of oral proficiency that it is difficult to find a currently used oral proficiency scale that explicitly refers to pronunciation. For all intents and purposes, the term accentedness has been replaced by comprehensibility and/or intelligibility. In comparison to sounding like a native speaker, intelligibility and comprehensibility are threshold-level criteria that represent achievable goals for L2 speakers to fulfill.

Derwing and Munro (2005) operationalize accentedness as strength of the perception of difference from a local norm (from no accent to a very strong accent), comprehensibility as listener processing ease (from extremely easy to impossible to understand), and intelligibility as "the extent to which a listener actually understands an utterance" (p. 385). Accentedness and comprehensibility are typically estimated through the use of 9-point Likert scales while intelligibility is usually associated with more explicit means of estimation, such as

percent/number of correct identifications of uttered words and phrases or performance on true/false and listening comprehension questions.

While highly influential, Derwing and Munro's operationalizations of comprehensibility and intelligibility have not led to consensus in the use of the terms nor in consensus on how they interact. Part of the problem lies in the overlap between the operationalizations of comprehensibility as the listener's ease of processing and intelligibility as the degree of actual comprehension. Derwing and Munro (2015) discuss the possible interactions between intelligibility and comprehensibility, as well as combinations between intelligibility and accentedness. They conclude that highly intelligible speech can be heavily accented. Highly intelligible may still require for a great amount listener effort to comprehend. All the possible combinations demonstrate the interaction within the two selected pairs of constructs: (a) intelligibility and comprehensibility, and (b) intelligibility and accentedness. As was summarize in Yan and Ginther (2017), listeners' individual perception of speakers' deviation from a norm influences their judgments about processing effort, as well as the amount of speech they are able to understand. Comprehensibility and intelligibility may be partially independent, but they also overlap. The use of the terms throughout the literature suggest that despite the extended efforts of Derwing and Munro, comprehensibility and intelligibility remain difficult to distinguish.

Perhaps the most influential contribution of Derwing and Munro's research has been to shift the instructional focus from the goal of achieving some facsimile of a native-like speech to a more realistic goal of accented but comprehensible and/or intelligible speech. while in agreement with the widely held skepticism concerning the native-like principle that "it is both possible and desirable to achieve native-like pronunciation in a foreign language" (Levis, 2005, p.370), Derwing and Munro do not recommend that pronunciation be abandoned. They reference

studies in which pronunciation instruction has been found to have a positive effect on both intelligibility and comprehensibility, and they argue that "prioritized pronunciation instruction" , or "a conceptualization of intelligibility that assists teachers in setting priorities" and the use of "empirical evidence that identifies effective practices" (Munro & Derwing, 2011) should focus on helping learners produce intelligible speech. Derwing and Munro (2009) explain:

> If time is spent on something that doesn't affect intelligibility or comprehensibility (such as the infamous interdental fricatives in English), something that really does matter will be neglected. Evidence is accumulating that what's important are the macroscopic things, including general speaking habits, volume, stress, rhythm, syllable structure and segmentals with a high functional load (Dewing & Munro, 2005). (pp. 482-483)

As Harmer (1991) states, "our aim should be to make sure that students can always be understood to say what they want to say. They will need good pronunciation for this, though they may not need to have perfect accents" (p. 22). Derwing and Munro's suggestion that instruction shift to intelligibility and comprehensibility to include features beyond segmental fidelity to native-like speech has been complemented by many studies that have examined the contributions of both segmental and suprasegmental aspects of speech. Also, the reduction of accentedness should not be a major issue or the primary issue at early stages of pronunciation instruction. Investigations of intelligibility, comprehensibility, fluency, in relation to speakers' overall oral language proficiency, support inclusion of an expanded set of variables— features of speech production that extend beyond segmentals. The next sections of literature review will discuss how accentedness if perceived in linguistics research, together with methodological concerns and solutions.

2.5.3. Perception of Accentedness

The perception of accentedness in speech has been examined through both segmental and suprasegmental features. The effects of segmental features on pronunciation perception are examined in combination with suprasegmental elements and fluency variables. Anderson-Hsieh, Johnson, and Koehler (1992) explored the relationship between native speakers' judgements of pronunciation in three areas: prosody, segmentals, and syllable structure. The strongest relationship was found between prosodic features and pronunciation ratings. Trofimovich and Baker (2006) investigated L2 acquisition of suprasegmentals by analyzing stress timing and tonal peak alignment in adult L2 speech, together with fluency measures of speech rate, pause frequency, and pause duration. Correlational results found speakers' approximation of English stress timing and fluency measures both impacted listeners' evaluation of speakers' accentedness. The predictive power of suprasegmental factors, however, was not as strong as that for fluency measures.

The contributions of segmental and suprasegmental features to accentedness and intelligibility, in particular, were also examined in Winters and O'Brien (2012). Intonation contours and syllable duration were mapped onto L2 speech and native speech of English and German. Listeners then completed cloze tests and comprehension tests to assess intelligibility. L2 segmental production was shown to have a stronger effect on accentedness perception. In addition, results for intelligibility tasks demonstrated an interaction between shared speaker and listener L1 background (Bent & Bradlow, 2003; Hahn, 2004). Listeners who share the same L1 background with speakers usually find the speech more intelligible.

Isaacs and Trofimovich (2012), examined listeners' comprehensibility ratings in relation to 19 quantitative speech measures represented by segmental, suprasegmental, fluency, lexical,

grammatical, and discourse-level variables which were then correlated with three L1 English listeners' scalar judgements of L2 speech comprehensibility. Correlational results found L2 comprehensibility ratings related to a wide range of variables not restricted to the domain of phonology and fluency. Reports from three experienced raters were collected for more detailed description of the features they found most noticeable. Their comments highlighted speakers' word stress, grammar, vocabulary, and fluency along with discourse structure and context representation. Raters' familiarity with the speakers' L1s was also mentioned as an influence on the comprehensibility ratings. While far from presenting the last word on accentedness, comprehensibility, and intelligibility, these studies provide a broader view of the components of pronunciation, including not only suprasegmental features, but also additional variables consisting of fluency and discourse features.

A developing line of speaking research examines the effects of different combinations of speakers' and listeners' L1s and L2s on listeners' perceptions of speaking performance. Reminiscent of Lado's (1964) call for comparisons between L1s and L2s to predict a learners' areas of difficulty, a growing number of studies have examined how listeners with different or the same language backgrounds process speech. Despite the well-received argument that strong accents do not necessarily lead to reported listener difficulty (e.g., comprehensibility as listener processing ease or difficulty), some studies report that L1 listeners appear to process L1 and L2 speech differently (Gibson et al., 2017), and that native speaker "disfluencies" (e.g., pausing) are strategic and may ease listener processing. As Bosker et al. (2014) comment, "It has been previously found that native disfluencies may help the listener in word recognition (Corley & Hartsuiker, 2011), in sentence integration (Corley, MacGregor, & Donaldson, 2007), and in reference resolution (Arnold et al., 2007)". (p. 609).

However, there appears to be, at least, an initial processing cost of accentedness for listeners. Ockey, Papageorgiou, and French (2016) and Ockey and French (2016) discuss performance effects on subjects' performance on listening comprehension items on a monologic task and then on an interactive lecture and found consistent debilitation of performance when even slightly accented speech was included in the input., Ockey and French (2016) explain that even accents judged to be light and completely comprehensible can influence test takers' performance on listening comprehension tasks based on interactive lectures. Strength of accent is a variable that needs to be carefully considered when considering the inclusion of accented speech in listening comprehension assessments.

However, studies examining the extent to which listeners adjust to speakers' accentedness have reported mixed but some positive, encouraging results for test takers. Gass and Veronis (1984) reported that listener processing costs diminish, often rapidly, with increased familiarity. Floccia, Goslin, Girard, and Konopczynski (2006) along with Clarke and Garrett (2004) also found evidence of initial processing difficulty followed by rapid processing adjustment/normalization (after as few as 2-4 utterances). Harding (2017) argues in a review of validity concerns for speaking assessments that the time has come for listener variables to be explicitly considered in construct definitions. Raters are invited to take a more active role in the drafting and application of speaking assessment rubrics, especially when a more clearly defined pronunciation scale is needed. Raters' familiarity and attitude towards speakers' pronunciation are also concerns for the generalizability inference in speaking assessment. In this dissertation, accentedness evaluation results come from the involvement of both speakers and listeners. Instead of focusing on only segmentals or suprasegmentals, speech samples selected for human rater assessment are generated based on fluency and lexica features, while speakers' L2 English

overall proficiency levels are controlled. The next section of literature review discusses how accentedness is evaluated by human raters, as well as how accentedness is operationalized based on previous definitions.

2.5.4. Developing Scales to Measure Accentedness

Interpreting accentedness as pronunciation errors or deviations in segmental and suprasegmental features has been challenged when addressing communication needs in broader social contexts. Following Derwing and Munro's arguments, Ockey and Wagner (2018) emphasized that instead of sounding like a native speaker, being comprehensible should be the primary goal of language learners. Accent is thus defined as "the way and degree to which a speaker's speech sounds different than the speech of speakers of the local variety (the speech variety to which the speaker's variety is compared)." (p.69).

Situated in an assessment context of higher level L2 English speakers, separating accentedness from intelligibility is a comparatively straightforward process. As for the Oral English Proficiency Test (OEPT), speakers who are rated above 50 must be intelligible (see OEPT scale in Appendix B). As the first rating descriptor on the OEPT holistic scale, being intelligible is the minimal prerequisite for speakers to fulfill before being rated above 50. The possible accentedness in their speech, however, provides an opportunity to explore its relationship with intelligibility and comprehensibility. Research on accentedness in the context of higher level L2 speakers is helpful in three aspects. First, the evaluation of accentedness needs to focus on "difference", or "variation from the local language norm", instead of involving intelligibility checks only, i.e. whether the speakers' verbal message can be understood. Second, documenting accentedness as a facet of linguistic profile, or a potential change in L2 speakers', may not be positively correlated with the growth of overall L2 proficiency. It is possible that

accent evaluation results for OEPT examinees rated as 60 will remain the same with those who are rated as 50. Thirdly, the reason for being accented is not confined to segmental or suprasegmental differences when higher level L2 English speakers are involved. Impact of other linguistic dimensions, such as speakers' choice of words and utterance fluency, may also contribute.

While trained and untrained human raters are asked to evaluate acentedness, 7-point or 9-point Likert scales are often used, which are discussed as interval data. Hayes-Harb (2014) mentioned methodological concerns regarding the research paradigm, who questioned the robustness of using Likert scales, the length of elicitation material, as well as the linguistic training the raters have received.

In order to address problems that accentedness possibly overlaps with intelligibility and comprehensibility, Ockey and French (2016) developed the Strength of Accent Scale (see Table 2.1). Listeners who used the scale were students and instructors from U.S. institutions with diverse disciplinary backgrounds and are either L1 English speakers or highly proficient L2 English speakers. Speakers evaluated by the Strength of Accent Scale, however, all have an L1 English background. The definition of accentedness emphasizes on perceived difference from a local variety and raters' judgment on comprehensibility. More justification for the scale was provided by Ockey (2018), who addressed three critical questions in the scale design: "(1) noticeability of differences in the speaker's speech from that of local variety, (2) effort required by the listener to accommodate to an accent, and (3) effort of difference in speech variety for understanding the message." (p. 86). The scale has been examined by Ockey (2018) to distinguish speakers' accents from a standard variety, where L2 listeners were reported to be slightly more severe in judging accentedness. Meanwhile, speech samples used in the validation

study were uniformly 20 seconds in length. The control of speech length helps eliminating possible interference of listeners' adjustment. Listeners may become familiar with the speaker's accent after longer periods of time, reporting that the speech samples are less difficult to comprehend.

Table 2.1 *Strength of Accent Scale (Ockey & French, 2016)*

Scale	Description
1	The speaker's accent is **NOT** noticeably different than what I am used to and **did NOT** require me to concentrate on listening any more than usual. The accent **did NOT** decrease my understanding.
2	The speaker's accent is **noticeably different** than what I was used to but **did NOT** require me to concentrate on listening any more than usual. The accent **did NOT** decrease my understanding.
3	The speaker's accent is **noticeably different** than what I was used to and **did** require me to concentrate on listening any more than usual. The accent **did NOT** decrease my understanding.
4	The speaker's accent is **noticeably different** than what I was used to and **did** require me to concentrate on listening any more than usual. The accent **slightly** decreased my understanding.
5	The speaker's accent is **noticeably different** than what I was used to and **did** require me to concentrate on listening any more than usual. The accent **substantially** decreased my understanding.

The Strength of Accent scale is adapted in this dissertation study based on situational contexts of the Oral English Proficiency Test. Trained OEPT raters primarily focus on intelligibility and are not asked explicitly to evaluate accent. They usually accommodate to examinees' accent during their rating practices and become lenient in deciding whether "the speaker's accent is different than what I was used to". In addition, advanced intermediate and advanced L2 English speakers do not cause comprehension difficulty based on OEPT holistic

scale descriptors. Level 3, 4, and 5 on the Strength of Accent scale may not be used at all, which opens up space for scale adaptation. The methods chapter of dissertation includes detailed steps adopted to modify the Strength of Accent Scale, where intelligibility is removed from the and listener effort is separated from accent difference.

2.6. Linguistic profile studies and cluster analysis in applied linguistics research

Linguistic profiling has been used in second language acquisition to describe the linguistic systems of groups of learners at a specific stage (Ågren, Grandfelt, & Schlyter, 2012; Bartning, 2000; Brindley, 1998; Clahsen, 1985; Pienemann, 1998, 2005; Pienemann & Keβler, 2011), or to compare individual sample with established profiles of different proficiency levels (Grandfeldt & Ågren, 2014; Keβler & Liebner, 2011). Learners display a variety of morphosyntactic and grammatical patterns of use at different proficiency levels, suggesting a developmental progression.

Linguistic profiling has also been explored through corpus-driven approaches with large amount of observational data. Research methods assume the correlations between semantic and distributional properties, or connections among distribution, form, and meaning (Divjak & Gries, 2006; Gries, 2010; Kuznetsova, 2015). Providing evidence for correlations between form and meaning, linguistic profiles generated from studies mentioned above can assist researchers with predicting meaning through the distribution of forms. In Russian language, for example, researchers can investigate "verbs that have a prevalence of masculine vs. feminine past tense endings in the corpus and examine the gender stereotypes that affect the activities denoted by the verbs." (Kuznetsova, 2015, p. 262).

Another strand of research investigates linguistic profiles through cluster analysis, a statistical method that recognizes homogeneity among data and places cases of similar numerical

attribution into the same group. Staples and Biber (2015) provide more details about the use of cluster analysis in the research of applied linguistics:

> Cluster analysis is a multivariate exploratory procedure that is used to group cases (e.g. participants or texts). Cluster analysis is useful in studies where there is extensive variation among the individual cases within predefined categories. For example, many researchers compare students across proficiency level categories, defined by their performance on a test or holistic ratings. But a researcher might later discover that there is an extensive variation among the students within those categories with respect to their use of linguistic features or with respect to attitudinal or motivational variables. (p. 243).

Cluster analysis helps extract evidence to analyze the characteristics of features used by L2 writers, or identify different types of L2 learners, which provides important information for educators to devise learning strategies. Callies, Diez-Bedmar, & Zaytseva (2014) retrieved and classified advanced L2 writers' use of reporting verbs. By using cluster analysis, advanced L2 writers were categorized into groups that use less diverse reporting verbs (e.g. say, state) to more diverse reporting verbs (e.g. discuss, argue). Researchers and educators can better visualize L2 writers' performance by observing certain linguistic features, clarify assessment objectives, and design developmental tasks. In Rysiewicz (2008), researcher used cluster analysis with middle school students' performance on language aptitude tasks and measures of L2 proficiency, presenting the cognitive profiles of successful and unsuccessful L2 English learners. Mechanical memory, defined as rote memory, did not play an important role in differentiating high and low achievers in L2 English learning. Inductive language learning abilities and expert use of first language, however, contributed to students' higher levels of achievement in learning English as a second language.

Cluster analysis has also been used to examine surface linguistic features of L2 English speakers' language test performance. Jarvis, Grant, Bikowski, and Ferris (2003) explored

multiple profiles of highly rated timed English compositions with 21 linguistic features, such as text length, conjuncts, hedges, and nominalization. The features included in the study cover general text characteristics, lexical features, as well as grammatical features. Research results show texts share similar within-group characteristics, but also demonstrate significant differences between groups**.** For example, certain clusters demonstrated a more frequent use of nouns than pronouns, while other clusters exhibit an opposite pattern. Frequent use of certain linguistic features varies across clusters, and the quality of writing depends on concerted application of different linguistic features. In other words, successful writers are better at devising strategies to compensate for deficiencies in their writing. Researchers also pinpointed some features (e.g. text length, lexical diversity, and conjuncts) that do not have a lot of variation across highly rated writings. Features such as mean word length and nominalization, however, differ across the profiles of highly rated essays as compared to those that were given lower holistic scores. Researchers thus stated that the quality of a written text may depend on how a collection of linguistic features in combination, instead of relying on the use of individual linguistic features.

Friginal, Li, and Weigle (2014) followed the models proposed by Jarvis et al. (2003) and used cluster analysis to identify linguistic profiles demonstrated in highly rated essays across native speaker (NS) and non-native speaker (NNS) groups. A number of 23 linguistic features were included for the cluster analysis, which established 6 profiles among all the NS and NNS writers. Most of the highly rated NNS essays were more formal and academic. In contrast, NS papers demonstrate a wide variety of different styles. Identification of profiles by using cluster analysis investigates the relationship between the distribution of linguistic features and general writing quality. Functional analysis of each profile and comparisons show that certain profiles

contain a predominant number of NS or NNS writers, suggesting teaching implications for writing instruction. Instructors can familiarize themselves with certain linguistic patterns across language learners' proficiency level and L1 background, which would help them select authentic texts as sample papers or provide students with individual feedback and guidance.

Cluster analysis, as a method rendering linguistic profiles, is also applied to explore the relationship between international students' English language proficiency and their college academic success. Ginther and Yan (2018) conducted cluster analysis on Chinese international students' TOEFL iBT subscale of listening, speaking, reading, and writing, where one of the clustered profiles had lower subscales in speaking and writing. A negative correlation between this particular group of students' TOEFL total score and their first-year grade point average was found, indicating that the relationship between academic success and language proficiency may display specific patterns within particular L2 English learner groups. Policy makers need to be aware that subscale test scores can produce different profiles, as international students' language test performance could be connected with their L2 English learning experience and strategies used for test preparation.

Measures to be included in the cluster analysis, however, are subject to the relationships among variables included in the analyses. Clustering results will be influenced when variables that are highly correlated with each other are included, resulting in collinearity and inaccurate profile extraction. More details about correlation reduction and profile selection will be discussed in the chapter of methods.

CHAPTER 3. METHODS

3.1. Research Question and Design

Focusing on English second language speakers of higher-level proficiency with a Hindi or Mandarin first language background, this dissertation examines speech samples rated as 50 and 60 for OEPT 2. Research questions are:

a) After cluster analysis, will selected variables for utterance fluency, lexical diversity, and lexical frequency help extract different linguistic profiles among high proficiency L2 English speakers?

b) What are the fluency and lexical variables that characterize profiles?

c) How will raters perceive the profiles in terms of accentedness and their efforts to comprehend the speakers?

3.2. Database Description

The Test of English as a Foreign Language (TOEFL) is composed of four sections: listening, speaking, reading, and writing. Each section has a subscale of 30 points. The cut point of 25 locates at the borderline between the Common European Framework of Reference (CEFR) proficiency level B2 and C1. CEFR has a scale range covering A1, A2, B1, B2, C1, and C2. While B2 is the advanced intermediate level where language users can follow academic instructions, C1 is the advanced level at which target language users can comfortably participate in activities such as teaching. Language users at the level of B2 will benefit from language instruction and support before handling teaching responsibilities. The Oral English Proficiency Test (OEPT) uses a five-level scale: 35, 40, 45, 50, and 55. The level of 50 is comparable to 25 on the TOEFL speaking subscale and corresponds to C1 on the CEFR scale.

This dissertation study examines test responses collected from the OEPT 2, which was administered from 2009 to 2015. A six-point scale was used (35, 40, 45, 50, 55, 60) for OEPT 2 L1 Hindi and L1 Mandarin speakers, who were rated 50 or 60 for the test.

For OEPT 2, examinees respond to three types of test items: generating opinions or offering advice with context information (Area of Study, Newspaper Headline, Compare and Contrast, Pros and Cons, Respond to Complaints), explaining graphs (Bar Graph, Line Graph), summarizing main ideas (Telephone Message, Conversation, Short Lecture), and reading short texts aloud (Read Aloud Text 1, Read Aloud Text 2). There are 12 test items in total. More information about OEPT 2 test items can be found in Appendix B and C. The Newspaper Headline item is the focus for this dissertation study. Examinees read a text prompt first, and then express their opinion as a member of the student and scholar community at this university. Short introductory texts are provided as background information. OEPT 2 has four test formats with four different Newspaper Headline prompts, which read as the following:

> **Form 1:** Do you think taking college courses online is a good way to study? Why or why not?
>
> **Form 2:** Do you think a television announcement will have a significant effect on the amount that students recycle? Why or why not?
>
> **From 3:** Do you believe that class size affects the quality of education? Why or why not?
>
> **Form 4:** Do you think it is the university's responsibility to prevent students from illegally downloading music? Why or why not?

Table 3.1 shows an inventory of transcribed OEPT 2 speech samples, which contains examinees' responses from Fall 2009 to Summer 2015. All of the 409 OEPT 2 speech samples were included in this study.

Table 3.1 *Transcribed Speech Samples from OEPT 2*

	Hindi 50	**Mandarin 50**	**Hindi 60**	**Mandarin 60**
Fall 2009	15	24	1	0
Spring 2010	4	9	0	0
Summer 2010	1	4	0	0
Fall 2010	11	34	1	0
Spring 2011	5	17	0	0
Summer 2011	2	4	0	0
Fall 2011	14	35	1	1
Spring 2012	1	19	1	0
Summer 2012	0	2	0	0
Fall 2012	1	38	10	3
Spring 2013	2	10	3	0
Summer 2013	0	3	0	0
Fall 2013	13	30	4	2
Spring 2014	2	13	0	2
Summer 2014	1	1	0	0
Fall 2014	4	29	7	2
Spring 2015	3	14	4	1
Summer 2015	1	0	0	0
Subtotal		**366**		**43**
Total	**80**	**286**	**32**	**11**

3.3. Research Phase I—Variable Coding

This dissertation study is divided into three phases, the first of which is a selection of representative measurement indices for two main constructs: utterance fluency and vocabulary. Five specific measures are coded and calculated for all the 409 speech samples. Table 3.2 presents a list of variables coded in this study: Mean Syllables per Run (MSR), Speech Rate

(SR), Pause Rate (PR), Measure of Textual Lexical Diversity (MTLD), and percentage of words on the Academic Word List (AWL). More detailed explanation of the five variables can be found in the literature review chapter, Appendix E, and Appendix F.

Table 3.2 *Measured Constructs and Coded Variables*

Construct	Interpretation	Measurement Index	Measurement Methods	Measurement Tool
Utterance Fluency	A combination of speed fluency and breakdown fluency	Mean Syllables per Run (MSR)	Mean Syllables per Run is calculated as the number of syllables divided by number of runs in a given speech sample. Runs are defined as numbers of syllables produced between two silent pauses. Silent pauses were considered as pauses equal to or no longer than 0.25 seconds (Ginther, Dimova & Yang, 2010).	Fluencing Software More information can be found in Park. S. (2016). *Measuring Fluency: Temporal Variables and Pausing Patterns in L2 English Speech*. (Unpublished Doctoral Dissertation). Purdue University, West Lafayette, IN.
	Speed Fluency	Speech Rate (SR)	Speech Rate (SR) is calculated as the number of syllables divided by response time.	
	Breakdown Fluency	Pause Rate (PR)	Pause Rate is defined as the number of filled and unfilled pause divided by response time.	

Table 3.2 continued

Construct	Interpretation	Measurement Index	Measurement Methods	Measurement Tool
Vocabulary Frequency	The frequency of vocabulary on the Academic Word List (AWL) in examinees' responses	Percentage of AWL words	The percentage words on the AWL list in each speaker's transcribed response.	AntWordProfiler: Available from http://www.laurenceanthony.net/software
Lexical Diversity	Lexical Diversity is explained as the variation of vocabulary used by speakers. More specifically, Lexical Diversity in this study is represented by the quantified relationship between type (class of words) and token (number of words).	Measure of Textual Lexical Diversity (MTLD)	Each word is evaluated sequentially for its TTR first. Factor count is conducted as the second step, which increases by 1 if a word has met the cutoff TTR value of 0.72. A partial factor count is also provided for the remainder of a lexical item, which is calculated as the range covered between 1.00 to 0.72. The ultimate calculation of MTLD result is fulfilled by having the total number of words divided by the total factor count.	Python program adapted from https://pypi.org/project/lexical-diversity/ based on McCarthy and Jarvis (2010). Step-by-step script is attached in Appendix A.

Mean Syllables per Run (MSR), Speech Rate (SP), and Pause Rate (RP) were calculated by *Fluencing*, a computer-assisted annotation tool introduced in Park (2016). Figure 3.1 shows a screenshot of the interface, which integrates transcription editor with audio player. Users mark overall pausing boundaries by observing the wave forms of each speech sample, dividing the audio files into shorter parts. After listening to each part, users are able to mark exact pausing boundaries with the actual speech sample transcription. Temporal and pausing information can be extracted and automatically calculated.

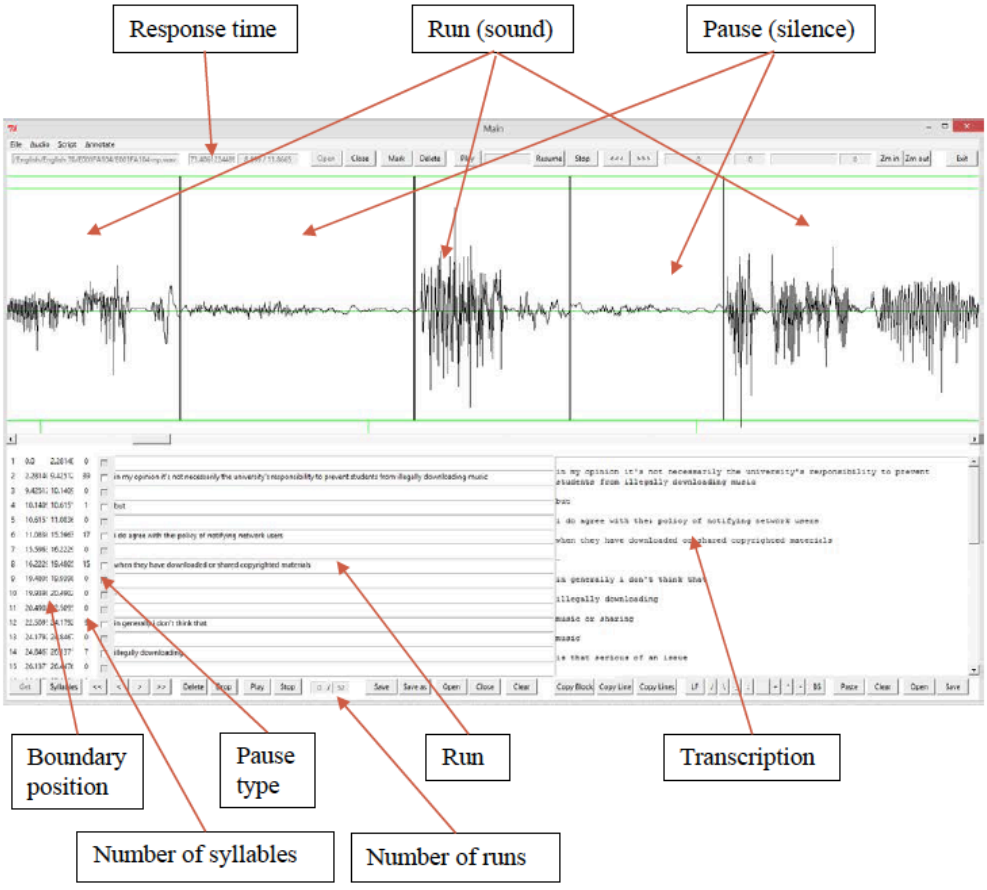


Figure 3.1 Sample Screenshot of the Fluencing Annotation Tool (Park, 2016: 46)

A space line is inserted between two consecutive speech runs during the transcription process. The researcher also used special characters to mark all the filled and non-filled pauses.

Table 3.3 *Special Characters Used for Transcribing Speech by Fluencing*

Special Character	Meaning	Example	Explanation
-	Filled pauses such as "er", "um","ah".	I do not think - taking online classes is a good idea	A filled pause occurred between speech runs "I do not think" and "taking online classes is a good idea". The filled pause is replaced with "-" and will not be considered as a syllable.
*	Syllables that could not be identified, but would be counted in the total number of syllables	I do not * taking online classes as a good idea	An unidentifiable syllable occurred at the end of "I do not", which is marked as "*". "*" will be counted as a syllable.

Three temporal variables were extracted from the *Fluencing* output: Mean Syllables per Run (MSR), Speech Rate (SR), and Pause Rate (PR). More information about the variable calculation methods can be found in Table 3.2. Mean Syllables per Run (MSR) is calculated by the total syllable number divided by the number of runs. Runs are defined as speech runs between two unfilled pauses longer than 0.25 seconds. Speech Rate (SR) is dividing the total number of syllables by response time, and Pause Rate (PR) is the total number of filled and unfilled pauses divided by response time. All the temporal variable results are saved in .json files, as is shown in Figure 3.2.

```
            "begin": 1808816,
            "end": 1815950,
            "syllables": 0,
            "tag": 0,
            "text": "-"
        },
        "84": {
            "begin": 1815950,
            "end": 1940480,
            "syllables": 6,
            "tag": 0,
            "text": "class size becomes really"
        }
    },
    "file_info": {
        "audio_name": "C:/Users/gao339/Desktop/Linguistic Profile/Chinese 60/FA-11-
Exam_8952/FA-11-8952-3-11619-np.wav",
        "frame_rates": 22050,
        "frames": 1841452,
        "total_time": 83.51256235827664
    },
    "speech_info": {
        "MSR": 8.588235294117647,
        "expected pausing rate": 1.0,
        "fp": 20,
        "pausing rate": 0.6106865669048122,
        "runs": 34,
        "sp": 31,
        "speech rates": 3.4964799516902967,
        "syllables": 292,
        "tags": 0
    }
}
```


Figure 3.2 Fluencing Output of Temporal Variables

Information about vocabulary frequency was retrieved from AntConc word profiler (Anthony, 2014). AntConc word profiler is loaded with three wordlists: GSL 1000 (the first 1000 words of the General Service List), GSL 2000 (the second 1000 words of the General Service List), and AWL (Academic Word List) 570 (Coxhead, 2000; West, 1953). Words of AWL are grouped as Level 3 in the output (Figure 3.4) and appear in blue within the transcribed text.

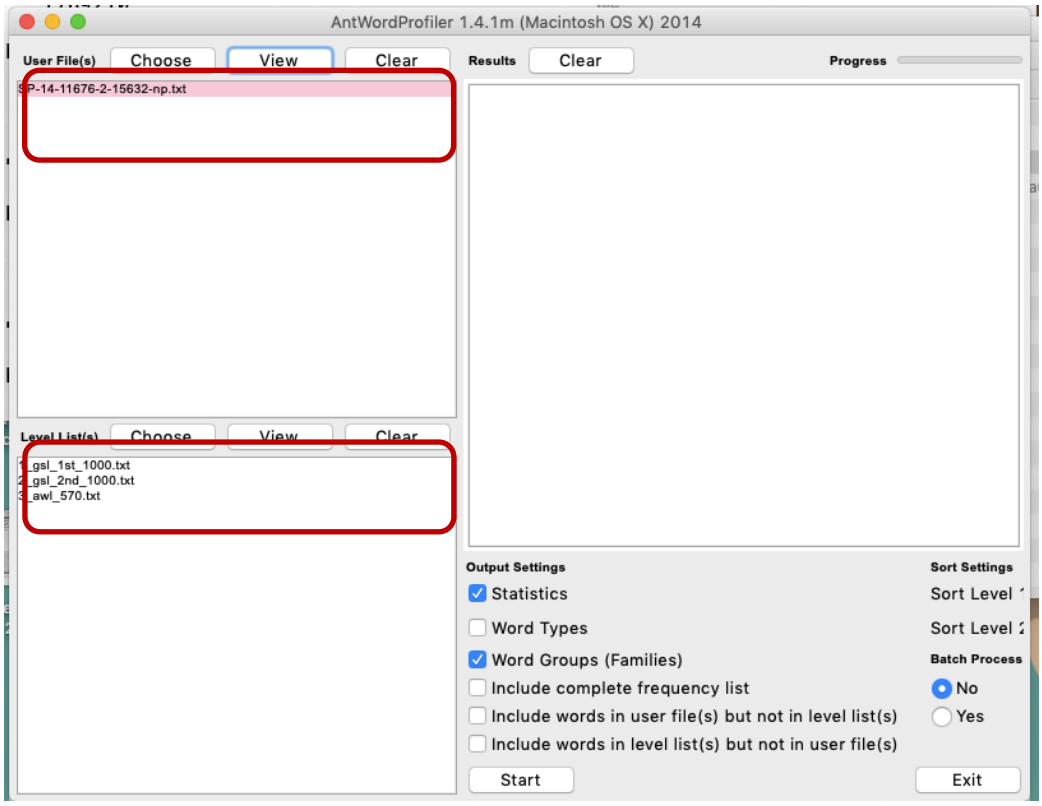


Figure 3.3 Interface of AntWord Profiler

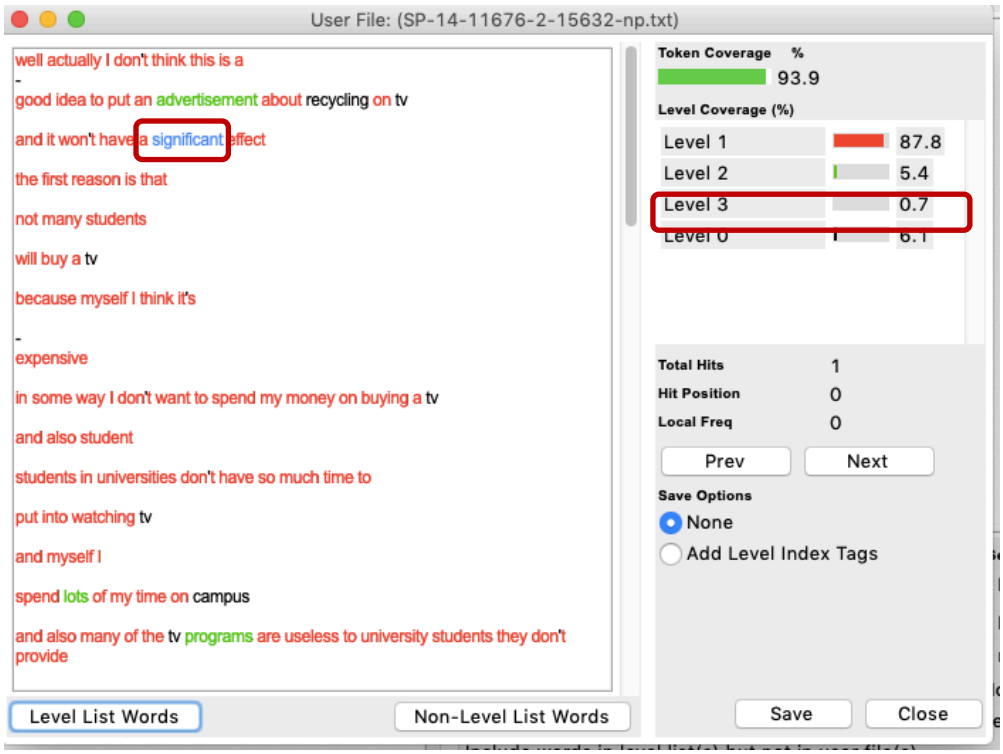


Figure 3.4 Vocabulary Frequency Information Retrieved from AntConc Word Profiler

As for the Measure of Textual Lexical Diversity (MLTD), open-source Python script is available at https://pypi.org/project/lexical-diversity/. The Python script is attached in Appendix A with adaptations made according to McCarthy and Jarvis (2010).

3.4. Research Phase II—Profile Identification

The second phase of this study is a Hierarchical Cluster Analysis (HCA) based on all the five fluency and vocabulary measures. As a method of classifying multivariate data into subgroups of homogeneity, cluster analysis helps extract linguistic profiles from all the four groups of speakers: L1 Hindi speakers rated as 50, L1 Mandarin speakers rated as 50, L1 Hindi speakers rated as 60, and L1 Mandarin speakers rated as 60.

Hierarchical Cluster Analysis (HCA) forms the backbone of cluster analysis (Everitt et al., 2010), where the concept of homogeneity and separation is of great importance. All agglomerative hierarchical methods ultimately reduce data into one single cluster, while divisive techniques help split data into difference groups. A series of data partitions are produced by agglomerative HCA. While the partition at the very end contains a number of speech samples, its counterpart at the highest level is inclusive of all the data cases. A certain number of clusters are obtained for all the 409 speech samples, which provide more detailed information about homogeneity and differences among groups of speakers.

Before conducting cluster analysis, correlational results between the variables need to be closely investigated. Due to the exploratory nature of cluster analysis, researchers may not be certain whether the variables selected are highly correlated. However, the correlation examination in this dissertation applies for this specific study only and bears limited inferential capacity, as data for each group of speakers are located within a restricted range of L2 English proficiency. No conclusion should be made as significant/insignificant correlation exists when a

group of L2 English speakers at a different proficiency level are involved. A Principal Components Analysis (PCA) may ensue to tackle the issue of correlation. If correlated variables load on the same dimension, the researcher could create a new index variable to be used for cluster analysis through linear combination, thus curbing the influence of collinearity and redundancy.

The main purpose of conducting Principal Component Analysis (PCA) is to optimally identify index variables from a larger set of measures. Figure 3.8 is an explanation of the working mechanism of PCA, where variable A_1, A_2, and A_3 are combined into one component C for further analysis. B_1, B_2, and B_3 are coefficient of the linear combination.

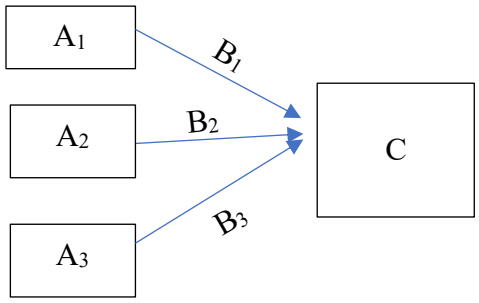


Figure 3.5 Principal Component Analysis

Phakiti (2018a) explicates the differences between Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA). Although both Exploratory Factor Analysis (EFA) and Principal Component Analysis (PCA) are exploratory in nature and can be used for dimension reduction, they differ in theoretical assumptions. EFA is grounded on the assumption that a latent variable explains all the observed variables. In PCA, however, the variances of observed variables are calculated to derive a component. As Phakiti (2018a: 424) concludes: "While EFA aims at generalizing to the target population, PCA only aims at reproducing the sample being used."

Before conducting PCA, all data need to be checked for two statistics: Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. KMO is a statistic that indicates the proportion of variance in variables that might be caused by underlying factors. High values close to 1 indicates that the sample size is adequate. If the value is less than 0.5, the results of the PCA would not be very useful. Bartlett's test of sphericity checks whether there is redundancy between variables that can be summarized with factors. Significance level value smaller than 0.05 suggests that PCA may be useful to further explanation of data. Components generated by PCA are to be used for cluster analysis as the next step.

After the Hierarchical Cluster Analysis (HCA), speakers' proficiency levels and L1 backgrounds were examined within each cluster. I made the hypothesis that speakers with a certain proficiency level tend to concentrate in a particular cluster, or clusters. In other words, speakers at the two proficiency levels should demonstrate different characteristics in their fluency and vocabulary features.

The purpose of conducting Hierarchical Cluster Analysis (HCA) is to place speech samples of each proficiency group into different clusters. An intuitive approach to select exemplary sample in each profile depends on the means of all the five fluency and vocabulary measures. I plan to select speech samples representative of each cluster for the next phase—accentedness evaluation from trained human raters. Together with results from Phase II, information about accentedness helps understand the characteristics of speakers from different L1 background.

3.5. Research Phase III—Accent Perception

Speech samples representing each profile were assigned to 12 trained raters for accent evaluation, who are familiar with the OEPT rating scale. All of the raters are either L1 English

speakers or highly proficient L2 English speakers, who hold a graduate degree in Teaching English to Speakers of Other Languages (TESOL) and are familiar with an array of English varieties. No raters have their first language background in Mandarin or Hindi, as common L1 background between listener and speaker may play a role in accent perception and the evaluation of listening comprehension efforts.

All of the speech samples are approximately 25 seconds in length to reduce possible listener accommodation effects. Ockey and French (2016) proposed that 20 seconds is a reasonable time span for listeners to evaluate speakers' accent. Given to the consideration that most speakers are able to finish one complete sentence during the 25-second time frame instead of 20 seconds, I edited the audio clips and used 25 seconds as a speech timing standard. As some speakers may use formulaic language or repeat test prompts at the beginning of their answer, I included the second sentence of speakers' response in edited speech samples.

Raters who participated in this study finished all the ratings online by using a Qualtrics survey link. Accent perception is explained through two sections: the difference between the speakers' accent and the local variety, and its possible influence on listeners' comprehension effort. Most audience members of international teaching assistants in the university community are domestic undergraduate students, who may not have been exposed to diversified English varieties and accents. The raters were thus informed that General American English is the baseline for this study. A short audio clip of an OEPT prompt is used as an example for General American English accent. Raters also went through a brief training session, where they used the scale to rate practice items. They were also asked to use the whole range of the scale. Six raters listened to the audio files in Order A as is shown in Table 3.4, where speakers' responses were

randomly arranged. To counterbalance possible order effect, the other six raters listened to the speech samples in Order B, which is the reversed version of Order A.

Table 3.4 *Rater Assignment for Speech Accent Evaluation*

	Speaker Number											
Order A	C	F	D	H	E	A	G	I	B	J	K	L
	Assigned to: Rater 1, Rater 2, Rater 3, Rater 4, Rater 5, Rater 6											
Order B	L	K	J	B	I	G	A	E	H	D	F	C
	Assigned to: Rater 7, Rater 8, Rater 9, Rater 10, Rater 11, Rater 12											

The two subscales from strength of accentedness scale developed by Ockey and French (2016) were adapted, as shown in Table 3.5.

Table 3.5 *Adapted Accentedness Measurement Scale from Ockey and French (2016)*

Part 1: How much is the accent different from what I am used to?

Scale	Description
1	The speaker's accent is almost the same with what I am used to.
2	The speaker's accent is slightly different than what I am used.
3	The speaker's accent is different than what I am used to.
4	The speaker's accent is noticeably different than what I am used to.

Part 2: How much listener effort is required?

Scale	Description
1	The speaker's accent did NOT require me to concentrate on listening any more than usual.
2	The speaker's accent requires me to concentrate on listening slightly more than usual.
3	The speaker's accent requires me to concentrate on listening more than usual.
4	The speaker's accent requires me to concentrate on listening much more than usual.

A partial credit Many-facet Rasch measurement (MFRM) model was used to analyze raters' evaluation of each speaker's accentedness and the efforts required to concentrate. MFRM models capture the influence of multiple variables on assessment outcome when a single scale is used. Eckes (2011) explains the application of MFRM in assessment situations where raters use one common rating scale. Three facets are identified in a case example: examinees, raters, and tasks. The MFRM model, which transforms observed ratings into a logit scale, is expressed by the equation in Figure 3.6.

$$In\left[\frac{p_{nljk}}{p_{nljk-1}}\right] = \theta_n - \sigma_l - \alpha_j - \tau_k$$

p_{nljk} = probability of examinee *n* receiving a rating of *k* from rater *j* on task *l*,

p_{nljk-1} = probability of examinee *n* receiving a rating of *k-1* from rater *j* on task *l*,

θ_n = proficiency of examinee *n*,

σ_l= difficulty of task *l*,

α_j= severity of rater *j*

τ_k = difficulty of receiving a rating of *k* relative to *k-1*

Figure 3.6 MFRM Model Equation (Eckes, 2011:14)

In this study, speakers, raters, and the two rating categories (accent difference and listener effort required) on the accent evaluation scale are the three facets included in the MFRM model. I decided to use a partial credit model, as the interaction between raters and the scale also needs to be considered. Table 3.6 is an excerpt of the rating data collected for MFRM analysis.

Table 3.6 *Excerpt Data Collection for Accent Evaluation*

				Criterion	
Cluster Membership	**L1 and Proficiency Level**	**Speaker**	**Rater**	**Accent Difference**	**Listener Effort**
1	Hindi 50	Speaker A	Rater 1	4	4
1	Mandarin 50	Speaker B	Rater 1	2	1
2	Hindi 50	Speaker C	Rater 1	4	2
2	Mandarin 50	Speaker D	Rater 1	3	2
3	Hindi 50	Speaker E	Rater 1	4	3
3	Hindi 60	Speaker F	Rater 1	4	3
3	Mandarin 50	Speaker G	Rater 1	4	3
3	Mandarin 60	Speaker H	Rater 1	2	1
4	Hindi 50	Speaker I	Rater 1	4	3
4	Hindi 60	Speaker J	Rater 1	2	1
4	Mandarin 50	Speaker K	Rater 1	4	2
4	Mandarin 60	Speaker L	Rater 1	2	1
….	….	….	….	….	….
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…		…	
4	Hindi 50	Speaker I	Rater 12	3	3
4	Hindi 60	Speaker J	Rater 12	2	2
4	Mandarin 50	Speaker K	Rater 12	3	1
4	Mandarin 60	Speaker L	Rater 12	2	3

CHAPTER 4. RESULTS AND DISCUSSION

4.1. Descriptive Statistics and Correlation Results

Descriptive statistics and box plots of the five measures of fluency and vocabulary are presented in Table 4.1 and Figure 4.2: Mean Syllables per Run (MSR), Speech Rate (SR), Pause Rate (PR), Measure of Textual Lexical Diversity (MTLD), and percentage of words on the Academic Word List (AWL).

Table 4.1 *Descriptive Statistics of Fluency and Vocabulary Measures*

	Variable	N	Mean	SD	Min	Max
Speakers rated as 50	MSR	366	7.63	1.85	3.47	17.62
	SR	366	188.72	27.24	107.4	282.00
	PR	366	.51	.11	.22	.86
	MTLD	366	46.53	12.32	24.1	110.4
	AWL	366	4.12	2.18	.00	14.30
Speakers rated as 60	MSR	43	10.5	2.69	6.47	17.70
	SR	43	222.49	32.07	124.2	276.6
	PR	43	.42	.09	.25	.70
	MTLD	43	52.15	13.47	26.61	91.19
	AWL	43	4.79	4.79	1.5	10.8

Boxplots in Figure 4.1 demonstrate that the five variables across the two proficiency levels are approximately normally distributed. The Kurtosis and Skewness statistics for all the variables are within the range between -.61 and 2.56. The Kurtosis is a statistic that examines

whether the data are heavily tailed towards an end, and the skewness checks the symmetry of a data set. Data with high Kurtosis (out of the range between -3 and 3) values are prone to have heavy tails or outliers. The assumption of normality is violated if the skewness statistics is over 3.

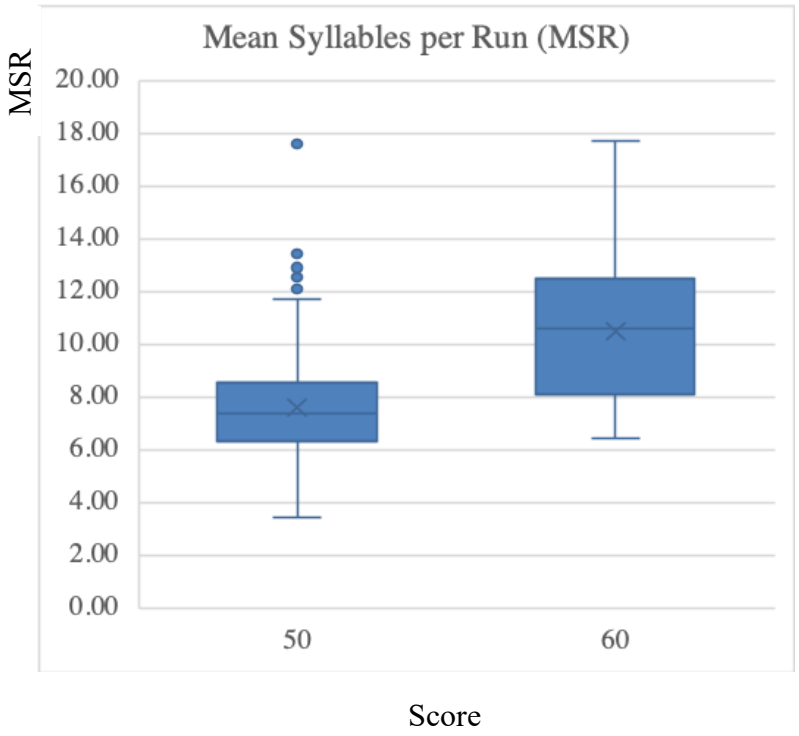
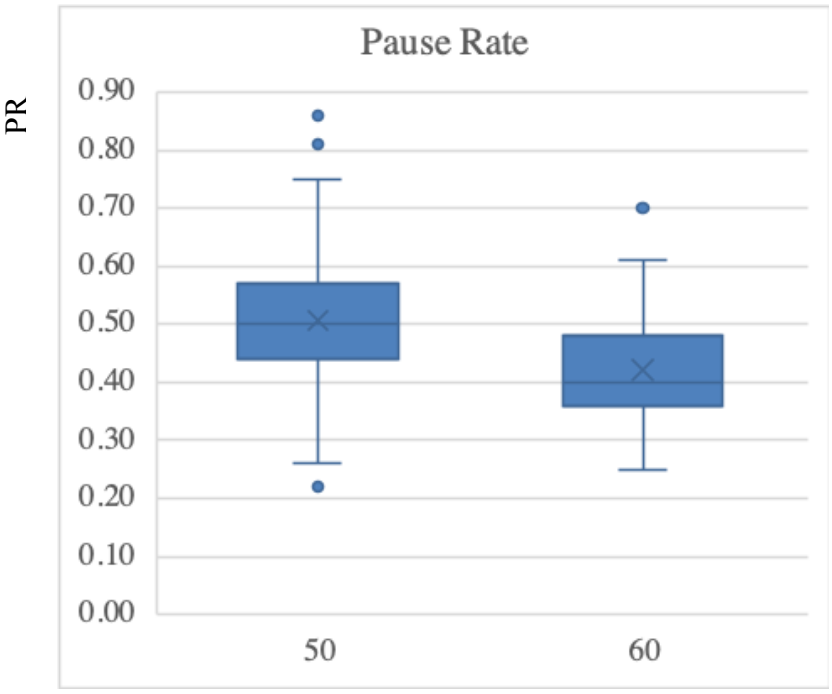


Figure 4.1 Boxplots of Fluency and Vocabulary Measures across Proficiency Levels

Figure 4.1 continued



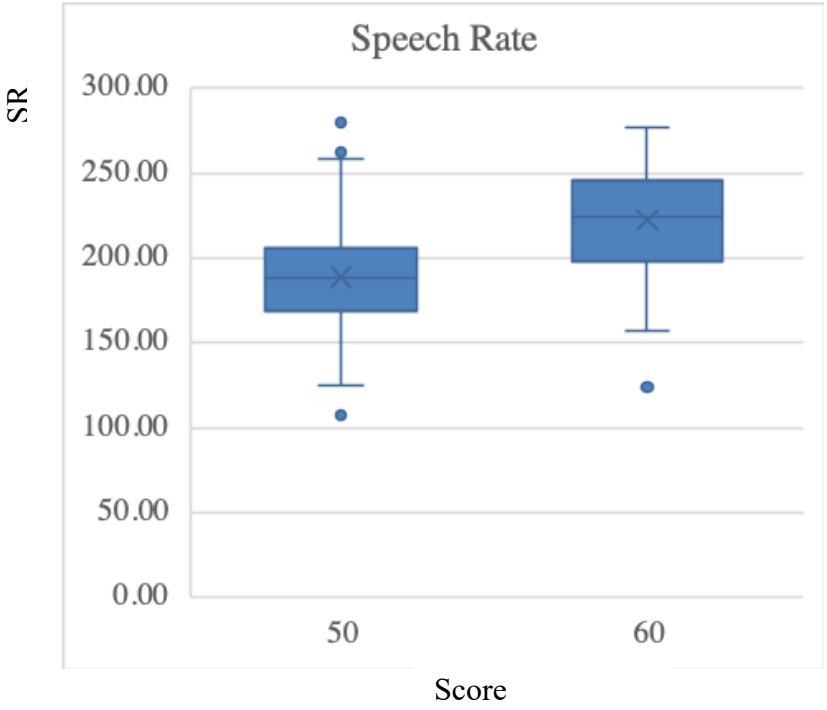
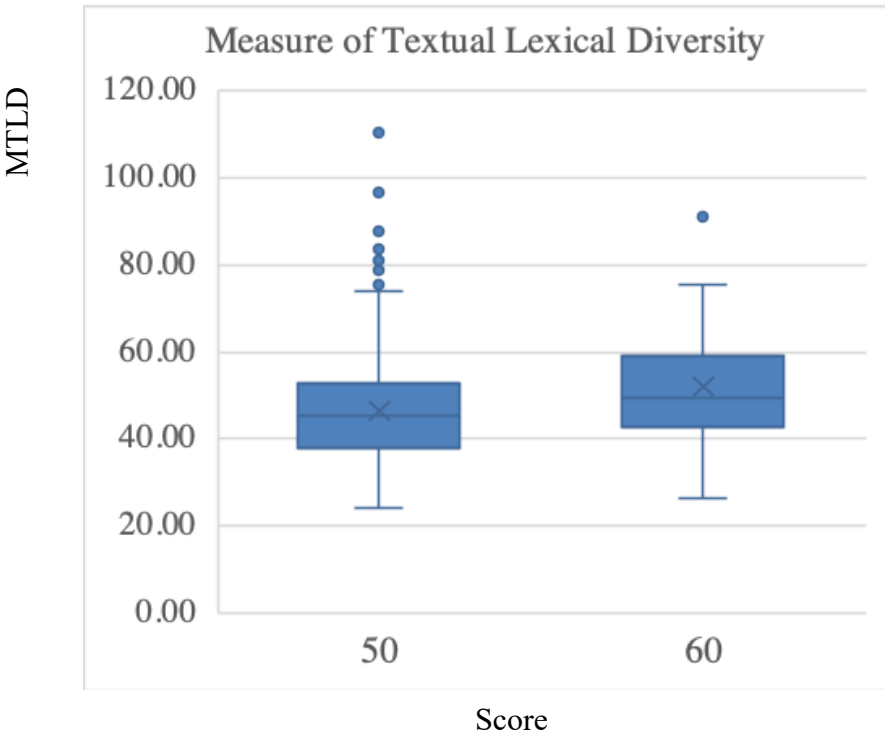
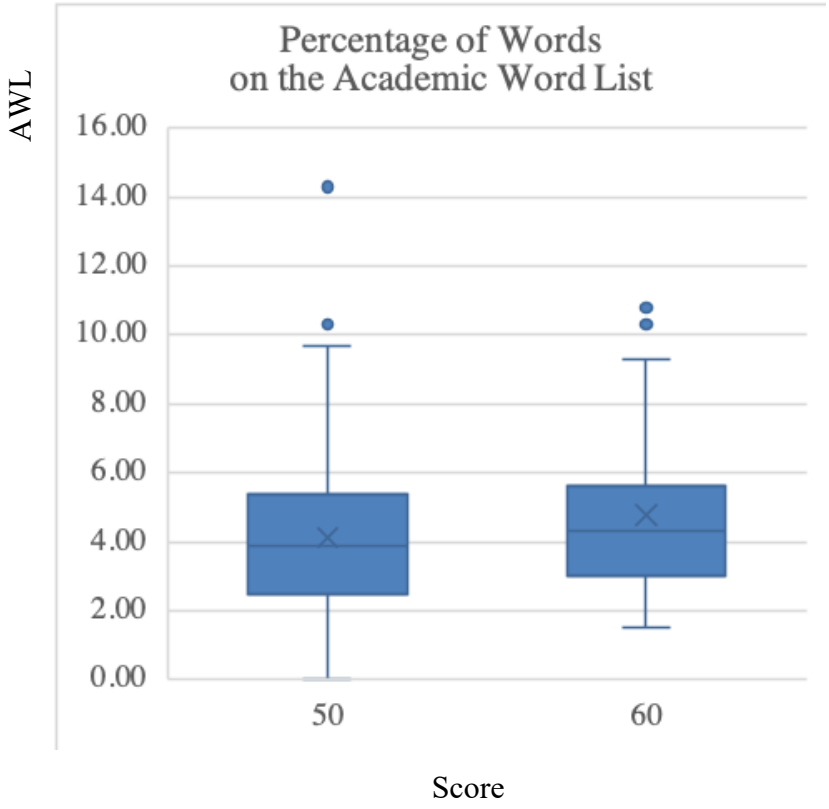


Figure 4.1 continued





Correlational results between variables are included in Table 4.2. For all of the speakers in this study, Mean Syllables per Run (MSR) is strongly correlated with Speech Rate (SR) and Pause Rate (PR). Pearson correlation values as 0.75 and -0.72 respectively. The results are not unexpected, as MSR is a composite variable that integrates both speed fluency and breakdown fluency. Keeping MSR in the study's variable collection is due to its strong effect in differentiating high proficiency L2 English speakers' performances. In comparison, correlations between fluency variables and vocabulary variables are moderate. The two vocabulary measures, Measure of Textual Lexical Diversity (MTLD) and the percentage of words on the Academic Word List (AWL), are correlated with each other to a lesser extent. Results suggest that Principal Component Analysis (PCA) is needed to reduce fluency and lexical variables, so that components to be used for cluster analysis will not cause collinearity issues. Two components are expected to be created after PCA, where the three fluency variables would load on one component and the two vocabulary features would load on another. The two new components will later be used for Hierarchical Cluster Analysis (HCA).

Table 4.2 *Correlation between Variables for All Speakers*

	MSR	SR	PR	MTLD	AWL
MSR	1				
SR	.75**	1			
PR	-.72**	-.41**	1		
MTLD	.11*	.12*	-.19*	1	
AWL	.18**	.12*	-.14**	.16*	1

** Correlation is significant at the .01 level (2-tailed)
* Correlation is significant at the .05 level (2-tailed)

4.2. Principal Component Analysis

Speakers rated as 50 and 60 were pooled together for Principal Component Analysis (PCA), so that common coefficients of linear combination could be obtained. For the pooled group of data, the KMO measure of sampling adequacy was close to 0.6, which is above the minimum value recommended for PCA. Bartlett's test of sphericity is less than 0.01. Oblique (Promax) rotation is used, as fluency and vocabulary measures are assumed to be related in explaining language proficiency test performance.

The scree plot in Figure 4.2 suggests that two components can be extracted. As is shown in Table 4.3, fluency measures are all significantly loaded on Component 1, while Component 2 includes the two vocabulary measures. Component 1 is thus named as fluency features, and Component 2 is named as vocabulary features. The two extracted components account for 68.68% of the variance among the five features, Correlation between the two components, which were used for the subsequent cluster analysis, is reduced to 0.20 after PCA.

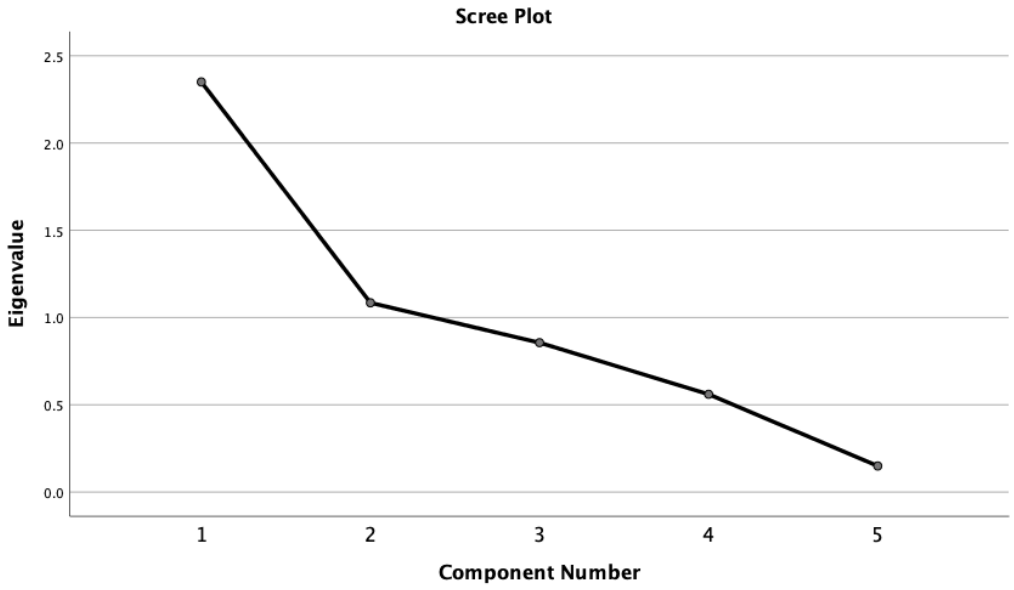


Figure 4.2 Scree Plot for Principal Component Analysis of Fluency and Vocabulary Measures

Table 4.3 *Component Loadings for Speakers Rated as 50 (after Promax Rotation)*

	Component 1 Fluency Features	**Component 2 Vocabulary Features**
Mean Syllables per Run (MSR)	.95	
Speech Rate (SR)	.84	
Pause Rate (PR)	-.80	
Measure of Textual Lexical Diversity (MTLD)		.80
Percentage of Words on the Academic Word List (AWL)		.70

4.3. Hierarchical Cluster Analysis Results

Hierarchical Cluster Analysis (HCA) was applied to data analysis with Ward's method of minimum within-group variance. I used two main techniques to decide the number of clusters in this study: a) Dendrogram observation and b) scree plot of coefficient change. Figure 4.3 shows the dendrogram generated for agglomerative HCA. The scree plot for coefficient change is presented in Figure 4.4.

Both the dendrogram (Figure 4.3) and scree plot (Figure 4.4) are references for deciding the number of clusters. The dendrogram in Figure 4.3 demonstrates a preliminary view of different clusters along the branches. The scree plot in 4.4 shows a bending point following a sharp decline of coefficients. Additional new cases are not creating new clusters after the bending point, which indicates that four-cluster solution is optimal in this case.

.

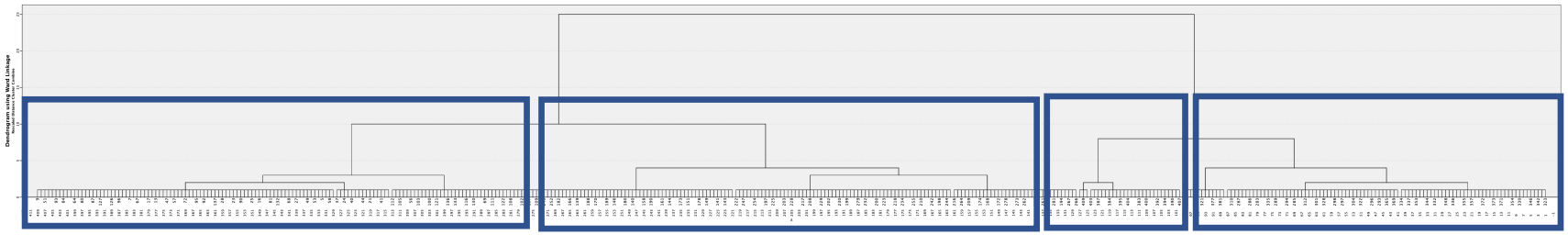


Figure 4.3 Hierarchical Cluster Analysis Dendrogram

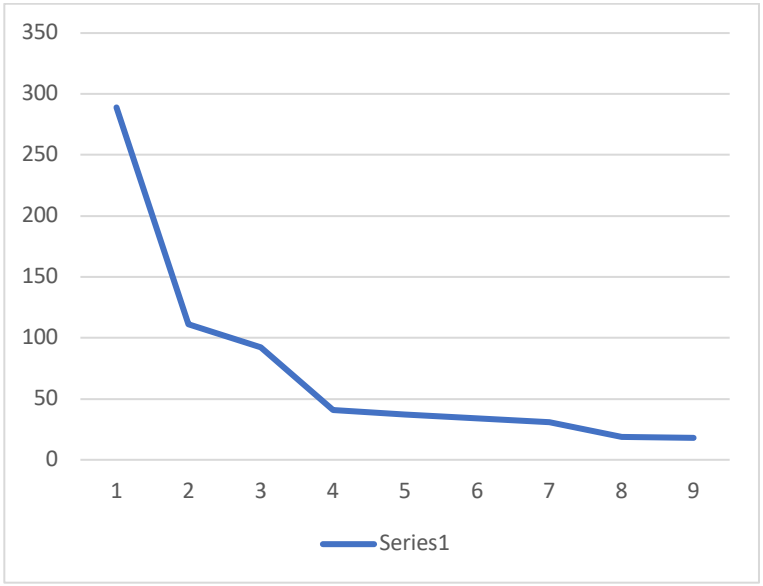


Figure 4.4 Hierarchical Cluster Analysis Scree Plot

Each of the four clusters generated from the Hierarchical Cluster Analysis (HCA) represents a profile. Before reporting the mean value of the five fluency and vocabulary features of each cluster, I transformed the cluster mean to an ordinal scale (Table 4.4).The ordinal scale, which reports the whole range of numerical values as a continuum, provides clearer comparison of variables among different profiles extracted from cluster analysis.

Table 4.4 *Four Levels to Describe the Five Features across the Two Proficiency Levels*

	Level of Description		
Mean Syllables per Run	Low	Medium	High
Speech Rate	Low	Medium	High
Pause Rate	Low	Medium	High
Measure of Textual Lexical Diversity	Low	Medium	High
Academic Word List	Low	Medium	High

In Table 4.5, I also hypothesized information of profiles extracted from HCA. Most of the speakers rated as 50 are in Cluster 1 and Cluster 2. Cluster 3 and Cluster 4 contain the majority of speakers rated as 60. Speakers were hypothesized to make progress in vocabulary features as their L2 English proficiency improves. Values of fluency features, however, do not always demonstrate a linear growth. For Cluster 3 and Cluster 4, whose members are mostly speakers rated as 60, some speakers may maintain the same level of fluency with speakers rated as 50.

Table 4.5 *Hypothesized Profile Information after Hierarchical Cluster Analysis (HCA)*

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Mean Syllables per Run (MSR)	Low	Medium	High	Medium
Speech Rate (SR)	Low	Medium	High	Medium
Pause Rate (PR)	High	Medium	Low	Low
Measure of Textual Lexical Diversity (MTLD)	Low	Low	Medium	High
Academic Word List (AWL)	Low	Low	Medium	High
	Speakers rated as 50		**Speakers rated as 60**	

Table 4.6 is the sample selection chart for accent evaluation. Two speech samples will be selected from each of the four clusters. One speaker has an L1 background in Hindi, and the other uses Mandarin as the L1.

Table 4.6 *Hypothesized Sample Selection for Accent Evaluation*

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Sample 1	Hindi 50	Hindi 50	Hindi 60	Hindi 60
Sample 2	Mandarin 50	Mandarin 50	Mandarin 60	Mandarin 60

Actual results from the cluster analysis are presented in Table 4.7, which includes descriptive statistics of Component 1 (fluency features) and Component 2 (vocabulary features) across clusters.

Table 4.7 *Descriptive Statistics of Component 1 (Fluency Features) Score and Component 2 (Vocabulary Features) Score across Clusters*

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Component 1 Fluency Features	-.80	-.13	1.17	.47	0
Component 2 Vocabulary Features	.22	-.84	.30	2.23	0

All the numerical variables have been transformed to ordinal scales. Table 4.8 lists more detailed information about the numerical range of each variable and the corresponding ordinal value. Table 4.9 demonstrates a closer examination at the five individual fluency and vocabulary variables, including mean values of each measure across the four clusters.

Table 4.8 *Ordinal Scale Conversion of Fluency and Vocabulary Features*

Ordinal Scale	Low	Medium	High	Very High
Mean Syllables per Run (MSR)	MSR < 7	7≤ MSR ≤9	9 < MSR ≤ 11	MSR> 11
Speech Rate (SR)	SR < 180	180 ≤SR ≤ 200	200 < SR ≤ 220	SR > 220
Pause Rate (PR)	PR ≤ 0.45	0.45 ≤ PR ≤ 0.50	0.50< PR ≤ 0.55	PR > 0.55
Measure of Textual Lexical Diversity (MTLD)	MTLD <45	45 ≤ MTLD ≤ 55	55 < MTLD≤65	MTLD >65
Academic Word List (AWL)	AWL <4	4≤ AWL ≤5	5 < AWL≤6	AWL > 6

In contrast to the hypothesis information listed in Table 4.5, Table 4.9 and Table 4.10 present the actual cluster analysis results. According to my expectations, speakers in Cluster 1 would demonstrate low values in both fluency features and vocabulary features. In comparison, speakers of Cluster 2 may show development in fluency measures but maintain the same level of vocabulary performance. However, results indicate that for speakers in Cluster 1, low utterance fluency is combined with medium values of Measure of Textual Lexical Diversity (MTLD) and percentage of vocabulary on the Academic Word List (AWL). Speakers of Cluster 2, who have medium fluency measure values, show lower values in both vocabulary features.

A similar situation also applies to Cluster 3 and Cluster 4. According to Table 4. I made the hypothesis that high fluency measures appear together with medium vocabulary measures, and medium fluency measures are combined high vocabulary measures. Final cluster analysis results in Table 4.9 and Table 4.10 show that speakers in Cluster 3, who have high utterance fluency measures, show medium or high vocabulary measurement results. Speakers in Cluster 4, who have medium utterance fluency measures, demonstrate very high vocabulary measurement results.

Table 4.9 *Descriptive Statistics of Fluency and Vocabulary Measures across Clusters*

	Number	**Mean MSR**	**Mean SR**	**Mean PR**	**Mean MTLD**	**Mean AWL**
Cluster 1	137	6.36 Low	170.03 Low	.56 Very High	50.63 Medium	4.25 Medium
Cluster 2	145	7.63 Medium	191.77 Medium	.52 High	38.02 Low	3.01 Low
Cluster 3	99	10.36 High	222.62 High	.40 Low	49.74 Medium	4.86 Medium
Cluster 4	28	8.63 Medium	199.91 Medium	.42 Low	69.61 Very High	7.67 Very High

Table 4.10 *Description of Mean Fluency and Vocabulary Measures for Each Cluster*

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Mean Syllables per Run (MSR)	Low	Medium	High	Medium
Speech Rate (SR)	Low	Medium	High	Medium
Pause Rate (PR)	Very High	High	Low	Low
Measure of Textual Lexical Diversity (MTLD)	Medium	Low	Medium	Very High
Academic Word List (AWL)	Medium	Low	Medium	Very High

The combination of fluency and vocabulary measures in Cluster 4 requires for more detailed examination. Values for fluency measures (Speech Rate and Mean Syllables per Run) in Cluster 4 are lower than Cluster 3 and are closer to those for Cluster 2. The vocabulary measures of Cluster 4, however, are noticeably higher than any other clusters. It is possible that speakers who use more diverse vocabulary and more academic words intended to control their delivery speed. Lower measures of Speech Rate (SR) and Mean Syllables per Run (MSR) could indicate higher proficiency in this occasion. This pattern may in turn have an effect on accentedness and effort ratings. There does appear to be an interaction, which will be discussed in section 4.5.

Further investigation into each cluster with Chi-square test shows that cluster membership is associated with both speakers' L1 background ($\chi^2 = 49.84$, $p < 0.01$) and overall oral proficiency level ($\chi^2 = 36.99$, $p < 0.01$). Table 4.11 lists the frequency number of speakers in each cluster based on their L1 background and OEPT 2 test scores. Most of the L1 Hindi speakers rated as 50 concentrated in Cluster 3. Cluster 3 also contains a large number of speakers rated as 60.

Table 4.11 *Cluster Membership Information*

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Total**
Hindi 50	22	16	32	10	80
Hindi 60	4	6	19	3	32
Mandarin 50	109	122	41	14	286
Mandarin 60	2	1	7	1	11
Total	137	145	99	28	409

The relationship between speakers' L1 background and their cluster membership is displayed in Figure 4.5, Figure 4.6, Figure 4.7, and Figure 4.8. Figure 4.5 is a percentage pie chart illustrating cluster membership of L1 Hindi speakers rated as 50: Among all the L1 Hindi speakers who were rated as 50, 27.5% of the speakers are in Cluster 1. 20% of the speakers are in Cluster 2. 40% of the speakers are in Cluster 3, and 12.5% of the speakers are in Cluster 4. The majority of L1 Hindi speakers are located in Cluster 3 based on the five fluency and vocabulary features.

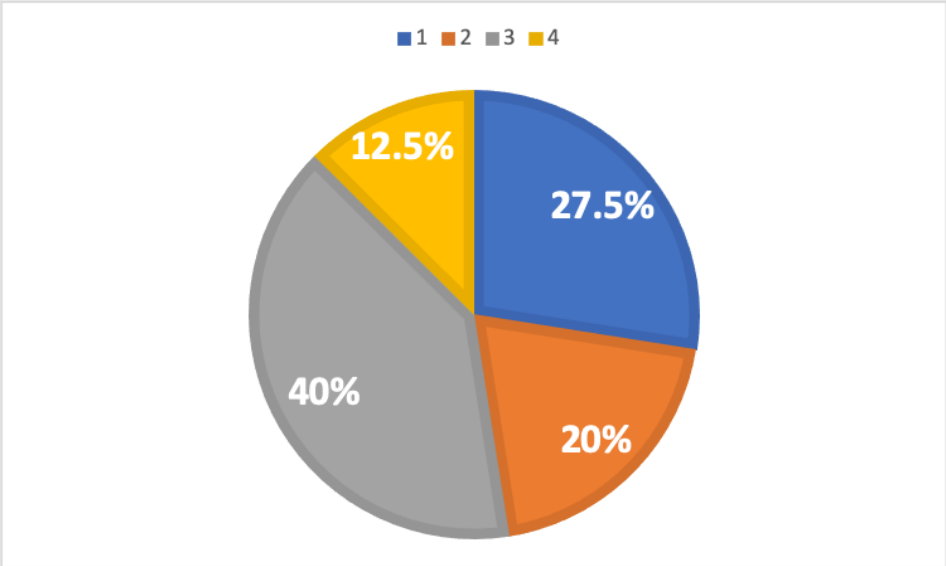


Figure 4.5 Cluster Membership of L1 Hindi Speakers Rated as 50

Figure 4.6 shows the percentage for L1 Mandarin speakers who were rated as 50. In comparison to L1 Hindi speakers, 38.11% percent of the L1 Mandarin speakers are in Cluster 1. 42.66% of the speakers are in Cluster 2. 14.34% of the speakers are in Cluster 3, and 4.9% of the speakers are in Cluster 4. In comparison to L1 Hindi speakers rated as 50, most of the L1 Mandarin speakers rated as 50 are in Cluster 1 and Cluster 2 instead of Cluster 3.

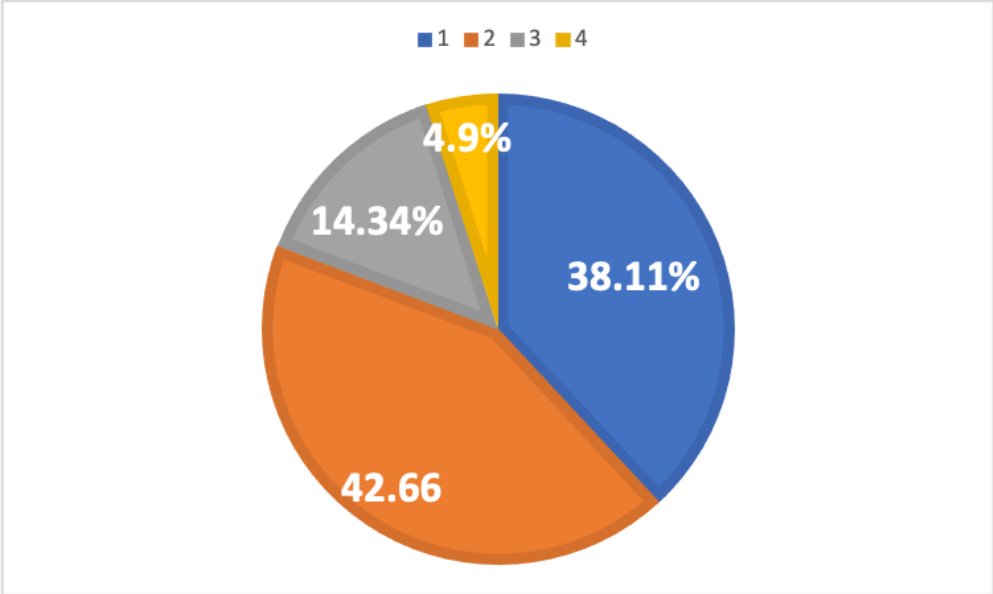


Figure 4.6 Cluster Membership of L1 Mandarin Speakers Rated as 50

The pattern for the two groups of speakers rated as 60, however, does not exhibit as great differences with the speakers rated as 50. As is shown in Figure 4.7 and Figure 4.8, most of the L1 Hindi speakers (59.38%) and L1 Mandarin speakers (63.55%) are in cluster 3. However, more L1 Hindi speakers (18.75%) are located in Cluster 2 when compared with L1 Mandarin speakers (9.08%).

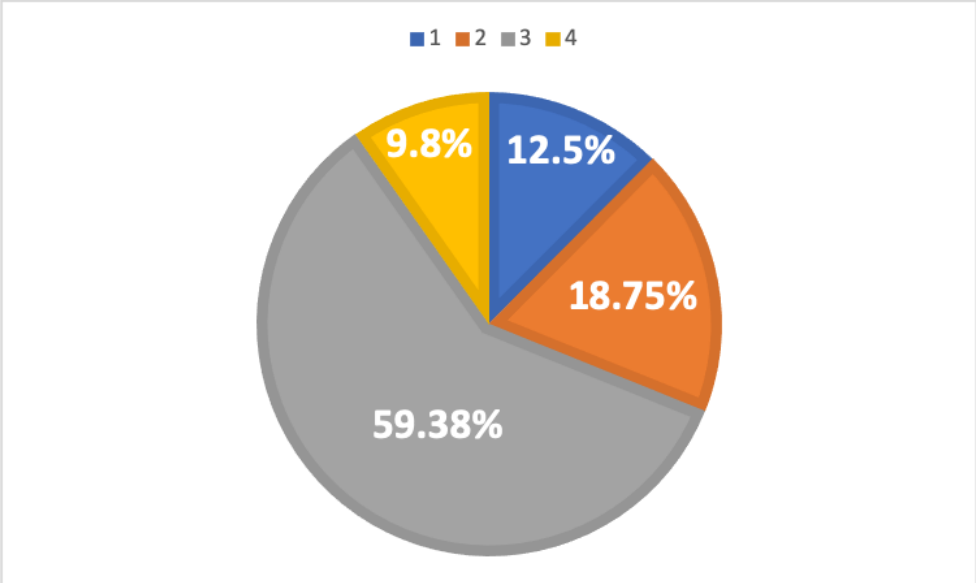


Figure 4.7 Cluster membership of L1 Hindi speakers rated as 60

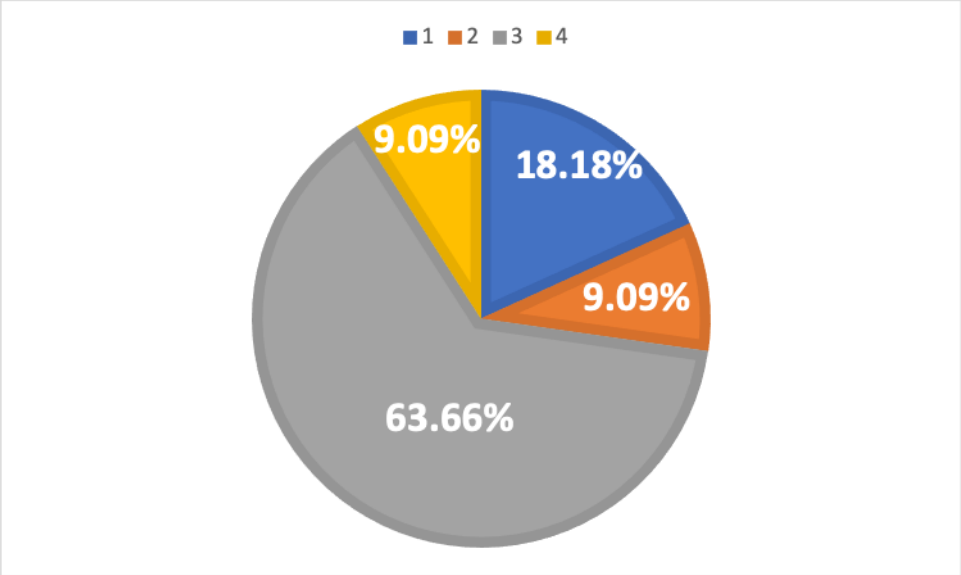


Figure 4.8 Cluster membership of L1 Mandarin speakers rated as 60

4.4. Sample Selection for Accent Evaluation

Table 4.12 exhibits the number of speakers selected from each cluster for accent evaluation. Based on the cluster membership information, most of the speakers rated as 50 are located in Cluster 1, Cluster 2, and Cluster 3. Samples for both L1 50 Hindi and L1 50 Mandarin

speakers are thus selected from the three clusters for accent evaluation. As most of the speakers rated as 60 are in Cluster 3, I have thus decided to select two speakers rated as 60 from Cluster 3 in addition to speakers rated as 50. The same decision was made for Cluster 4 as well. According to Table 4.10, speakers in cluster 4 have high performance values in both utterance fluency and vocabulary features. It would be beneficial to include all the four group of speakers in Cluster 4 to detect possible variations in accentedness.

All of the speech samples selected from each cluster have their feature values closest to the cluster mean, which are listed in Table 4.13. Only a small percentage of L1 Mandarin speakers were rated as 60 in OEPT, leading to the situation that Cluster 3 and Cluster 4 have only one L1 Mandarin speaker rated 60. However, fluency and vocabulary measures of their speech sample are not far from the cluster mean. The two selected L1 Mandarin speakers rated 60 are thus considered to be representative of Cluster 3 and Cluster 4.

Table 4.12 *Speech Sample Selection Chart*

Cluster Membership	**Speakers' L1 Background and Overall Proficiency Level**			
Cluster 1	Hindi 50		Mandarin 50	
Cluster 2	Hindi 50		Mandarin 50	
Cluster 3	Hindi 50	Hindi 60	Mandarin 50	Mandarin 50
Cluster 4	Hindi 50	Hindi 60	Mandarin 60	Mandarin 60

Table 4.13 *Information of Selected Speech Samples*

Cluster Membership	Speaker No.	L1 Background and Proficiency Level	MSR	SR	PR	MTLD	AWL
Cluster 1	Speaker A	L1 Hindi 50	6.44	154.80	0.60	48.01	3.50
	Speaker B	L1 Mandarin 50	6.28	160.20	0.44	50.18	3.70
Cluster 2	Speaker C	L1 Hindi 50	7.67	213.00	0.44	37.68	3.90
	Speaker D	L1 Mandarin 50	7.60	75.20	0.49	36.93	1.30
Cluster 3	Speaker E	L1 Hindi 50	10.44	225.60	0.38	48.49	4.10
	Speaker F	L1 Hindi 60	10.55	242.40	0.42	42.62	4.80
	Speaker G	L1 Mandarin 50	10.18	201.00	0.35	51.05	6.20
	Speaker H	L1 Mandarin 60	9.37	232.20	0.46	45.24	5.40
Cluster 4	Speaker I	L1 Hindi 50	9.19	213.00	0.41	78.68	5.80
	Speaker J	L1 Hindi 60	8.15	198.00	0.39	59.00	10.60
	Speaker K	L1 Mandarin 50	8.46	77.60	0.36	66.28	4.50
	Speaker L	L1 Mandarin 60	9.88	233.40	0.49	75.52	6.30

4.5. Accent Evaluation Results

A total number of 144 ratings were collected from 12 raters. Descriptive statistics in Table 4.14 show that a whole range of scale points from 1 to 4 have been used. Skewness and Kurtosis statistics for both the two items on the scale, i.e., the difference between speaker's accent and the local variety, as well as listeners' efforts required to concentrate, are within the range between -1.05 to 0.23. The results indicate that the ratings are not heavily tailed towards either end. In other words, raters have used a range of scale scores to differentiate speakers' accent variation and their efforts in listening. The Cronbach's alpha value reaches 0.90 between the two items, which indicates that accent difference and listener effort are strongly related.

Table 4.14 *Descriptive Statistics of Accent Evaluation Results*

Rating Scale	N	Mean	SD	Range	Skewness	Kurtosis
Accent Difference	144	2.49	1.00	3.00	.02	-1.04
Listener Effort	144	2.18	0.95	3.00	.23	-1.05

A more detailed analysis includes frequency bar graphs of each speaker and Many-Facet Rasch Measurement (MFRM) of raters' use of scale. Frequency results of ratings for each speaker present preliminary patterns in raters' judgement. MFRM takes individual raters' understanding of the scale into consideration and transforms speakers' accent evaluation results into logit scales. The logistic transformation results were added to the finalized profile information. Importantly this transformation forces the data into a normal distribution

Before extracting exact values about accent difference and effort difference from MRFM analysis, I include frequency bar graph in Figure 4.9, Figure 4.10, Figure 4.11, and Figure 4.12 to

demonstrate raters' perceptions of all the 12 speakers. Raters need to evaluate the difference between the speaker's accent and the General American accent, as well as the effort expended. The rating scale for both the two categories range from 1 to 4, with lower scores indicating less accent difference and lower listener effort.

L1 Hindi speakers and L1 Mandarin speakers in Cluster 1 and Cluster 3 show noticeable differences in raters' accent evaluation. L1 Mandarin speakers were consistently rated as less accented. Compared with L1 Hindi speakers. L1 Mandarin speakers also required less effort. In Cluster 2 and Cluster 4, however, evaluation results for accents and listener effort are not as obvious. Further MRFM analysis is thus needed.

Figure 4.9 illustrates raters' evaluation for Speaker A (L1 Hindi speaker rated as 50) and Speaker B (L1 Mandarin speaker rated as 50) in Cluster 1. Their speaking performances demonstrate low Speech Rate (SR), low Mean Syllables per Run (MSR), very high Pause Rate (PR), medium level of Measure of Textual Lexical Diversity (MTLD), and medium percentage of vocabulary on the Academic Word List (AWL). Speaker A (L1 Hindi speaker rated as 50) has most of the ratings located in 3 and 4 for raters' evaluation of accent difference and listener effort while Speaker B (L1 Mandarin speaker rated as 50) are rated predominantly with 1 and 2.

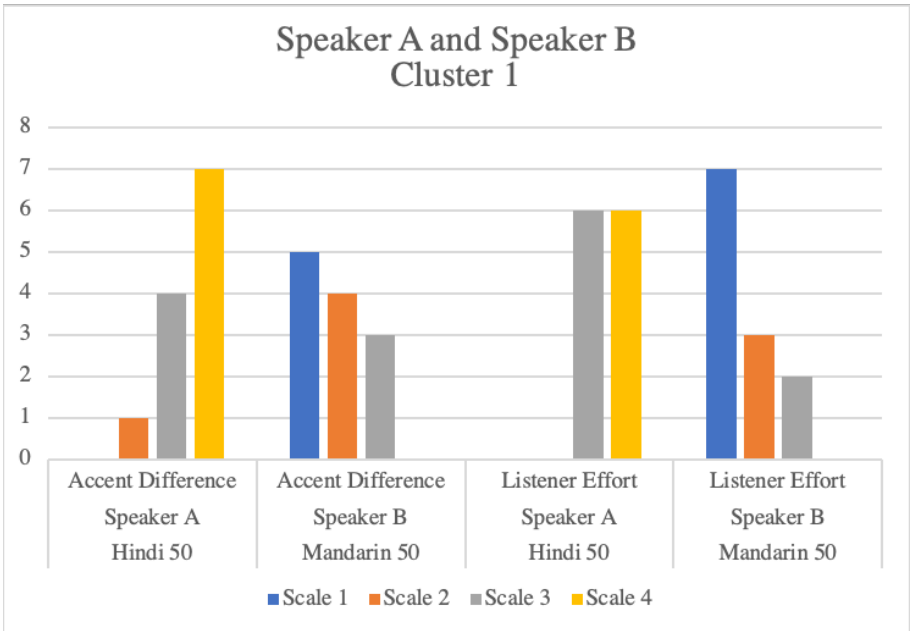


Figure 4.9 Accent Evaluation Results for Cluster 1

The same trend of accent evaluation occurred to speakers in Cluster 3. Figure 4.10 presents rating results for the four speakers in Cluster 3: Speaker E (L1 Hindi speaker rated as 50), Speaker F (L1 Hindi speaker rated as 60), Speaker G (L1 Mandarin Speaker rated as 50), and Speaker H (L1 Mandarin speaker rated as 60). Speakers in Cluster 3 demonstrate high Speech Rate (SR), high Mean Syllables per Run (MSR), low Pause Rate (PR), medium Measure of Textual Lexical Diversity (MTLD), and medium percentage level of vocabulary on the Academic Word List (AWL).

Speaker E and Speaker F within Cluster 3, who both have L1 background of Hindi, do not demonstrate a clear difference in accent change when the proficiency level changes from 50 to 60, Also, more listener effort is also associated with Speaker E and Speaker F. For the two L1 Mandarin speakers, however, Speaker H (rated as 60) has the most ratings of 1 in both accent difference and listener effort. Compared with Speaker E (L1 Hindi speaker rated as 50), Speaker

G (L1 Mandarin speaker rated as 50) has more ratings in 1 than 2. Raters tend to find that Speaker G is not as accented as Speaker E. They also spent less effort to concentrate when Speaker G was responding.

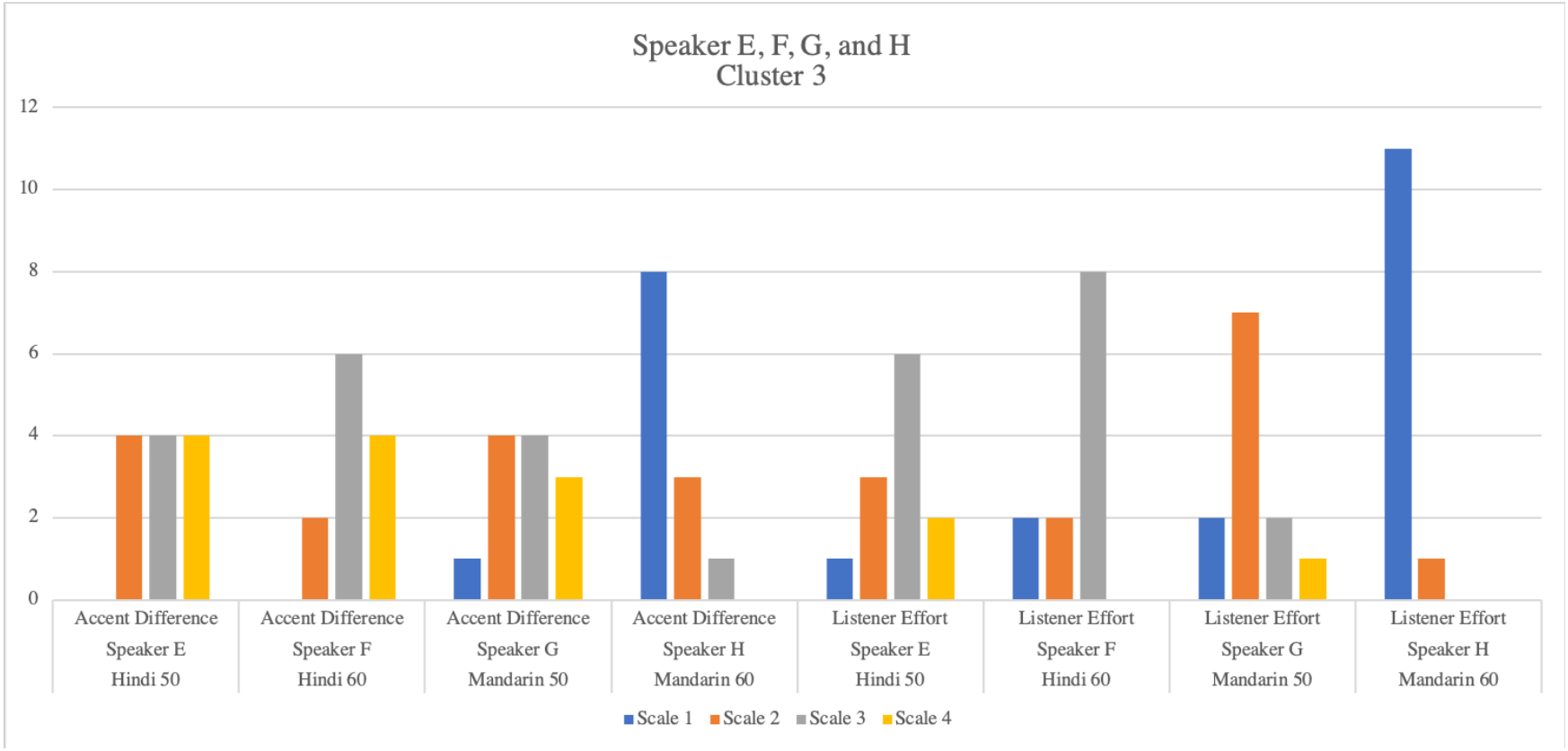


Figure 4.10 Accent Evaluation Results for Cluster 3

L1 Hindi Speakers and L1 Mandarin Speakers in Cluster 2 and Cluster 4, however, do not demonstrate clear-cut differences. Figure 4.11 shows rating results for Speaker C (L1 Hindi speaker rated as 50) and Speaker D (L1 Mandarin speaker rated as 50) in Cluster 2. Their speaking performances demonstrate medium Speech Rate (SR), medium Mean Syllables per Run (MSR), high Pause Rate (PR), low level of Measure of Textual Lexical Diversity (MTLD), and low percentage level of vocabulary on the Academic Word List (AWL).

In comparison to Cluster 1, which also includes a great number of speakers rated as 50, accent evaluation results for the two speakers in Cluster 2 do not show a clear difference based on the frequency bar graph. Both Speaker C and Speaker D have more ratings at scale level 2 or 3. Raters indicate that the two speakers' accents are slightly different or different from the General American English variety and they are required to concentrate slightly more or more while listening.

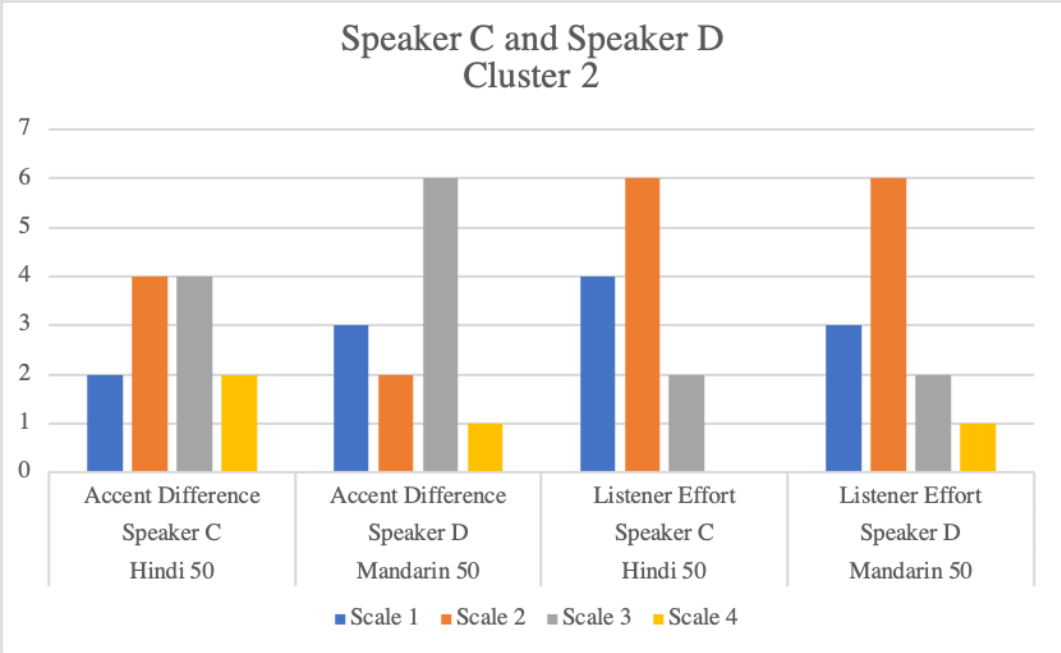


Figure 4.11 Accent Evaluation Results for Cluster 2

Accent evaluation results for Cluster 4 requires for further investigation. Speakers of Cluster 4 demonstrated medium Speech Rate (SR), medium Mean Syllables per Run (MSR), low Pause Rate (PR), very high Measure of Textual Lexical Diversity (MTLD), and very high percentage level of vocabulary on the Academic Word List (AWL). Speakers in Cluster 4 have similar fluency measures with Cluster 2, but have the highest measures of vocabulary among all the 4 clusters. Figure 4.12 shows the rating results for the four speakers in Cluster 4: Speaker I (L1 Hindi speaker rated as 50), Speaker J (L1 Hindi speaker rated as 60), Speaker K (L1 Mandarin speaker rated as 50), and Speaker L (L1 Mandarin Speaker rated as 60).

Speaker J (L1 Hindi speaker rated as 60) and Speaker L (L1 Mandarin speaker rated as 60) have lower ratings in both accent difference and listener effort. Although ratings for accent difference are close between Speaker J (L1 Hindi speaker rated as 60) and Speaker L (L1 Mandarin speaker rated as 60), Speaker J (L1 Hindi speaker rated as 60) has lower ratings in listener effort than Speaker L (L1 Mandarin speaker rated as 60). Speaker J (L1 Hindi speaker rated as 60)'s low ratings in accent difference and listener effort may attribute to his/her accommodation in slowing down the delivery speed.

Results for speakers rated as 50, however, are more straightforward. Speaker K (L1 Mandarin speaker rated as 50) is less accented than Speaker I (L1 Hindi speaker rated as 50), and raters also reported less listener effort when evaluating the speech sample of Speaker K (L1 Mandarin speaker rated as 50).

The many-facet Rasch measurement (MFRM) analysis helps pinpoint specific logit scale of each speaker for their accent evaluation results. The MFRM analysis in this study includes three facets: speaker, rater, and two items (accent difference and listener effort) for accent evaluation. Figure 4.13 is the wright map showing performances of the three facets.

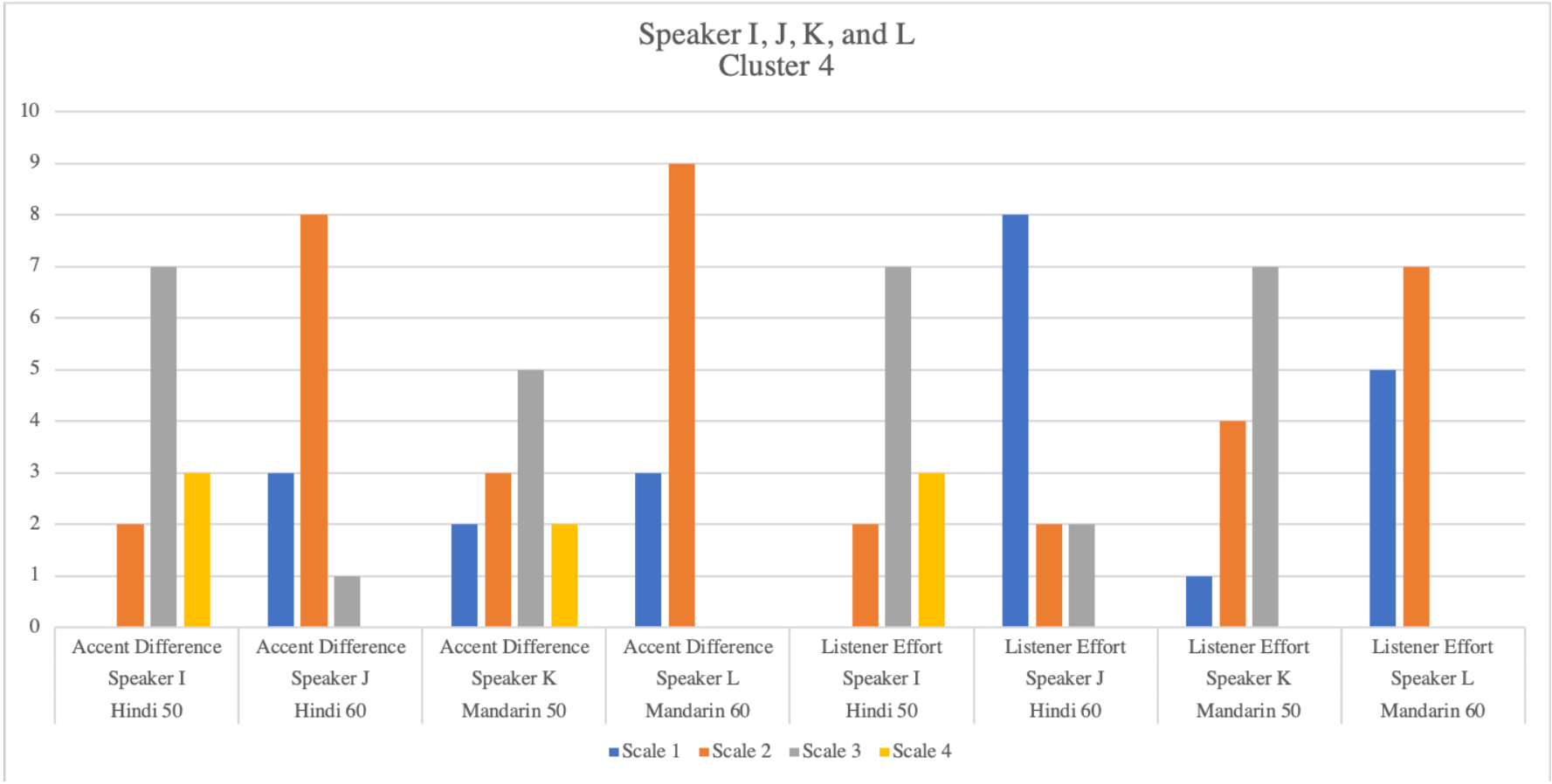


Figure 4.12 Accent Evaluation Results for Cluster 4

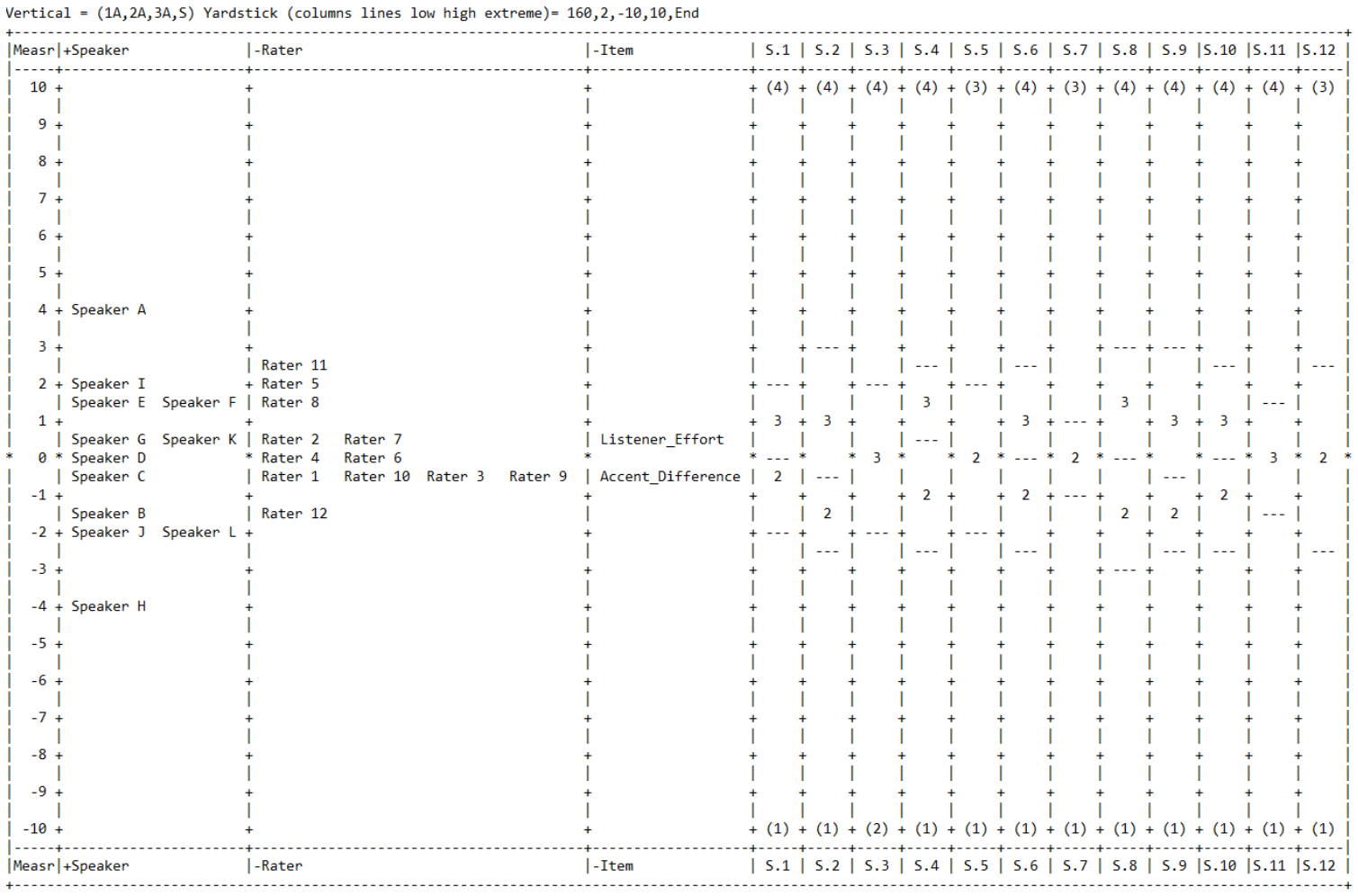


Figure 4.13 Wright Map for MFRM Results

Logit scales for L1 Hindi speakers and L2 Mandarin speakers are included in Table 4.16 and Table 4.17. As for MRFM results, higher logit numbers indicated stronger accent and a larger amount of listener effort required. The logit value for L1 Hindi speakers are in the range of -1.86 to 3.78, whereas the range for L1 Mandarin speakers is between -3.89 to 0.49. Logit scales were also transformed to ordinal scales (Table 4.15), which provide more information for speakers' performance within each profile.

Table 4.15 *Ordinal Scale for Accent Evaluation*

Range	**Scale**
0 and Below	Low
0-0.5	Medium
0.5-1	High
1 and Above	Very High

Table 4.16 *Logit Scale for L1 Hindi Speakers*

	L1 Hindi Speakers					
	Speaker A Cluster 1	Speaker C Cluster 2	Speaker E Cluster 3	Speaker F Cluster 3	Speaker I Cluster 4	Speaker J Cluster 4
OEPT Score	50	50	50	60	50	60
Cluster Membership	1	2	3	3	4	4
Logit Score	3.78	-0.30	1.40	1.28	2.03	-1.86
	Very High	Low	Very High	Very High	Very High	Low

Table 4.17 *Logit Scale for L1 Mandarin Speakers*

	L1 Mandarin Speakers					
	Speaker B Cluster 1	Speaker D Cluster 2	Speaker G Cluster 3	Speaker H Cluster 3	Speaker K Cluster 4	Speaker L Cluster 4
OEPT Score	50	50	50	60	50	60
Cluster Membership	1	2	3	3	4	4
Logit Score	-1.71	-0.07	0.49	-3.89	0.44	-1.86
	Low	Low	Medium	Low	Medium	Low

L1 Mandarin speakers were generally rated as less accented and required for less listener effort. According to the ordinal scales, four L1 Hindi speakers (Speaker A, Speaker E, Speaker F, Speaker I) scored very high in accent evaluation. The other two L1 Hindi speakers (Speaker C and Speaker J) were rated as low. Four of the six L1 Mandarin speakers (Speaker B, Speaker D, Speaker H, and Speaker L) were rated as low in terms of accent difference and listener effort required. The other two L1 Mandarin speakers (Speaker G and Speaker K) have medium ratings in terms of accent evaluation.

Table 4.18 presents research results for this dissertation study. In Cluster 1 and Cluster 2, L1 Mandarin speakers (Speaker B and Speaker D) have low ratings in accent difference and listener effort required. However, L1 Hindi speaker (Speaker A and Speaker C) have higher logit numbers in both the two categories.

Across Cluster 3 and Cluster 4, the two L1 Mandarin speakers rated as 50 (Speaker G and Speaker K) scored as medium in accent evaluation, while the L1 Mandarin speakers rated as 60 (Speaker H and Speaker L) were rated as low. L1 Mandarin speakers experienced a drop in accentedness along with the growth of their overall L2 English oral proficiency. Similarly, L1 Hindi speakers rated as 60 in both Cluster 3 and Cluster 4 (Speaker F and Speaker J) have lower ratings in accent evaluation than L1 Hindi speakers rated as 50 (Speaker E and Speaker I). In

conclusion, both groups of L1 Hindi and L1 Mandarin speakers sound less accented along with the growth of overall L2 English oral proficiency.

While the accent evaluation results remain medium or low across Cluster 3 and Cluster 4 for L1 Mandarin speakers, L2 Hindi speakers showcase different combinations of linguistic features. For L1 Hindi speakers rated as 50, Speaker I in Cluster 4 with slower delivery speed and higher vocabulary measures was rated more accented than Speaker E in Cluster 3. For L1 Hindi speakers rated as 60, however, Speaker F in Cluster 3 with faster delivery speed and lower vocabulary measures was evaluated to be more accented than Speaker J in Cluster 4.

Table 4.18 *Profile Information for all the Speakers across Cluster*

	Cluster 1		**Cluster 2**		**Cluster 3**				**Cluster 4**			
	Hindi 50	**Mandarin 50**	**Hindi 50**	**Mandarin 50**	**Hindi 50**	**Hindi 60**	**Mandarin 50**	**Mandarin 60**	**Hindi 50**	**Hindi 60**	**Mandarin 50**	**Mandarin 60**
Mean Syllables per Run (MSR)	Low	Low	Medium	Medium	High	High	High	High	Medium	Medium	Medium	Medium
Speech Rate (SR)	Low	Low	Medium	Medium	High	High	High	High	Medium	Medium	Medium	Medium
Pause Rate (PR)	Very High	Very High	High	High	Low	Low	Low	Low	Low	Low	Low	Low
Measure of Textual Lexical Diversity (MTLD)	Medium	Medium	Low	Low	Medium	Medium	Medium	Medium	Very High	Very High	Very High	Very High
Academic Word List (AWL)	Medium	Medium	Low	Low	Medium	Medium	Medium	Medium	Very High	Very High	Very High	Very High
Accent Difference from Local Variety	Very High	Low	Low	Low	Very High	High	Medium	Low	Very High	Low	Medium	Low
Listener effort	Very High	Low	Low	Low	Very High	Very High	Medium	Low	Very High	Low	Medium	Low
	Speaker A	**Speaker B**	**Speaker C**	**Speaker D**	**Speaker E**	**Speaker F**	**Speaker G**	**Speaker H**	**Speaker I**	**Speaker J**	**Speaker K**	**Speaker L**

CHAPTER 5. CONCLUSION AND IMPLICATION

5.1. Profile Information of Advanced Intermediate and Advanced L2 English Speakers

This dissertation study investigated linguistic profiles of 409 advanced intermediate and advanced L2 English speakers with two different L1 backgrounds: Mandarin Chinese and Hindi. All the L2 English speakers were international graduate students, who were administered the Oral English Proficiency Test (OEPT) at Purdue University. Three fluency-rated variables (Mean Syllables per Run, Speech Rate, and Pause Rate) and two vocabulary-related variables (Measure of Textual Lexical Diversity and percentage of words on the Academic Word List) were measured. Principal Component Analysis (PCA) shows that the three fluency-related variables load on one component, and the two vocabulary-related variables load on another component. These two components were used for Hierarchical Cluster Analysis (HCA), which extracts four different profiles among all the L2 English speakers. The four linguistic profiles demonstrate different combinations of the five linguistic features.

Profile 1: Low Mean Syllables per Run (MSR), low Speech Rate (SR), very high Pause Rate (PR), medium Measure of Textual Lexical Diversity, and medium percentage of words on the Academic Word List (AWL)

Profile 2: Medium Mean Syllables per Run (MSR), medium Speech Rate (SR), high Pause Rate (PR), low Measure of Textual Lexical Diversity, and low percentage of words on the Academic Word List (AWL)

Profile 3: High Mean Syllables per Run (MSR), high Speech Rate (SR), low Pause Rate (PR), medium Measure of Textual Lexical Diversity, and medium percentage of words on the Academic Word List (AWL)

Profile 4: Medium Mean Syllables per Run (MSR), medium Speech Rate (SR), low Pause Rate (PR), very high Measure of Textual Lexical Diversity, and very high percentage level of words on the Academic Word List (AWL).

Twelve experienced ESL instructors listened to the responses representing each linguistic profile first, and then evaluated speech accentedness and required listener efforts. As most of the speakers rated as 50 are located in Cluster 1 and Cluster 2, I selected one L1 Hindi speaker rated as 50 and one L1 Mandarin speaker rated as 50 for Cluster 1 and Cluster 2. Two speakers rated as 50 from each L1 background were also selected from Cluster 3 and Cluster 4. In addition to speakers rated as 50, speakers rated as 60 were also selected from Cluster 3 and Cluster 4, which resulted in a total number of 12 selected speakers from the four linguistic profiles. All of the selected responses have values of the fluency and vocabulary features closest to the cluster mean.

Frequency counts and multi-facet Rasch measurement (MFRM) analysis show that for Cluster 1, Cluster 3, and Cluster 4, advanced intermediate L1 Hindi 50s were additionally rated higher on accentedness and listener effort than L1 Mandarin 50s. In Cluster 3 and Cluster 4, both L1 Hindi speakers and L1 Mandarin speakers experienced a drop in accentedness evaluation as their OEPT scores increase from 50 to 60. The accent drop in Cluster 3 for L1 Hindi speakers, however, is not as prominent compared with Cluster 4.

A noticeable observation from the research results is the trade-off effect between fluency and vocabulary features. Speakers in Cluster 1 demonstrate medium values in fluency features and low values in vocabulary features. The situation is reverse for Cluster 2, where speakers have enhanced values of vocabulary measures but still deliver at a lower speed. Similar trade-off phenomenon can also be found in Cluster 3 and Cluster 4. Speakers of Cluster 3 demonstrated high values in fluency features, in combination with medium values in vocabulary features. Speakers in Cluster 4, however, have medium-level fluency features but very high values in vocabulary features.

5.2. Connections among L2 speakers' Overall L2 Proficiency, L1 Background, Accentedness, Fluency, and Vocabulary Features

Speakers' accentedness evaluation results are connected with both their L1 background and proficiency levels. As I did not examine the segmentals and supresegmentals of all the speech samples, the evaluation of accentedness largely depends on raters' perception. L1 Mandarin speakers tend to have lower ratings in accentedness and effort. In Cluster 3 and Cluster 4, which include speakers at both advanced intermediate and advanced L2 English oral proficiency levels, L1 Mandarin speakers and L1 Hindi speakers rated as 60 were evaluated to be less accented than advanced intermediate speakers rated as 50.

The impact of fluency and vocabulary features in combination on accent evaluation, however, does not work the same for speakers with different L1 backgrounds across L2 English oral proficiency levels. As L1 Hindi speakers rated as 50, Speaker A and Speaker I have the highest ratings in accentedness and effort (3.78 and 2.03 as logit scales respectively). Compared with Speaker C and Speaker E (L1 Mandarin speakers rated as 50), who received lower ratings in accentedness and listener effort, Speaker A and Speaker I both demonstrate slower delivery speed, higher Measure of Textual Lexical Diversity, and higher percentage of academic words.

Advanced L1 Hindi speakers rated as 60 display a different pattern. Speaker F (logit scale 1.28) in Cluster 3 is rated as more accented and requires more listener effort than Speaker J (logit scale -1.86) in Cluster 4. Speaker F in Cluster 3 delivers at a faster speed and uses less diverse vocabulary or academic words than Speaker J in Cluster 4. Speaker J (L1 Hindi speaker rated as 60), who is evaluated to be the least accented L1 Hindi speaker with the highest overall L2 English oral proficiency, may have slowed down delivery speed while using diverse vocabulary and a great percentage of words on the Academic Word List (AWL).

L2 English oral proficiency level does not have a huge influence on accent evaluation for Speaker E (L1 Hindi speaker rated as 50) and Speaker F (L1 Hindi speaker rated as 60), both of whom are in Cluster 3. Although Speaker F has a lower logit scale in accent evaluation (1.28) than Speaker E (1.40), the difference is minor. Both the two speakers located within the range of "very high" when logit scales revert to ordinal values. It may be the case that L1 Hindi speakers who have high delivery speed and use vocabulary of medium diversity are generally rated as highly accented

As for the L1 Mandarin speakers, L2 English oral proficiency level plays an important role in raters' evaluation of accentedness. Speaker H in Cluster 3 and Speaker L in Cluster 4, who are both advanced L2 English speakers with an L1 background of Mandarin, were evaluated to be less accented and require less listener effort. For advanced L1 Mandarin speakers rated as 60, Speaker L (logit scale -1.86) in Cluster 4 who delivers with lower values in fluency features and has higher vocabulary features sounds more accented than Speaker H (logit value -3.89) in Cluster 3. This situation is to the contrary of advanced L1 Hindi speakers rated as 60.

For advanced intermediate L1 Mandarin speakers rated as 50, Speaker D (logit scale as -0.07) in Cluster 2 and Speaker G (logit scale as 0.49) in Cluster 3 have higher ratings in accentedness evaluation and listener effort required than Speaker B (logit value 1.71) in Cluster 1 and Speaker K (logit value 0.44) in Cluster 4. Dissimilar with advanced intermediate L1 Hindi speakers rated as 50, Speaker B and Speaker K, who were evaluated with a stronger accent, demonstrate higher delivery speed, lower Measure of Textual Lexical Diversity, and uses fewer academic words.

In conclusion, the combinations of fluency, vocabulary, and accentedness are reverse from L1 Hindi to L1 Mandarin speakers. Advanced intermediate L1 Hindi speakers with the

combination of lower fluency measures and higher vocabulary measures (Speaker A in Cluster 1 and Speaker I in Cluster 4) were rated as more accented, while advanced intermediate L1 Mandarin speakers with the combination of higher fluency measures and lower vocabulary measures tend to be given higher ratings for accentedness and effort (Speaker D in Cluster 2 and Speaker G in Cluster 3).

The situation flips when advanced L2 English speakers are involved. Advanced L1 Hindi speaker with the combination of higher fluency measures and lower vocabulary measures (Speaker F) received higher ratings in accent, while advanced L1 Mandarin speaker with the combination of lower fluency measures and lower vocabulary measures (Speaker L) was evaluated to be more accented.

Accent evaluation might be connected with different accommodation strategies adopted by advanced L2 English speakers. Speaker J (L1 Hindi speaker rated as 60) and Speaker H (L1 Mandarin speaker rated as 60) have the lowest logit scale on accent evaluation among all the 12 speakers. Speaker J (L1 Hindi speaker rated as 60) in Cluster 4 was evaluated to be the least accented L1 Hindi speaker, who speaks slower and uses more diverse vocabulary and more words on the Academic Word List. Speaker H (L1 Mandarin speaker rated as 60) in Cluster 3 was evaluated to be the least accented L1 Mandarin speaker, who speaks faster and uses less diverse vocabulary and fewer words and Academic Word List. Given to the strong association between accentedness and listener effort required, advanced L2 English speakers may adjust their strategies while fine-tuning their delivery skills. While L1 Hindi speakers may attempt to slow down their delivery speed and diversify vocabulary usage, L1 Mandarin speakers can speed up and use less complicated vocabulary.

5.3. The distribution of Advanced Intermediate and Advanced L2 English Speakers in the Four Clusters

Chi-square tests show that cluster membership is strongly associated with speakers' L1 background ($\chi^2 = 49.84$, df = 3, $p < 0.01$) and overall oral proficiency level ($\chi^2 = 36.99$, df = 3, $p < 0.01$). Most of the advanced intermediate L1 Hindi speakers rated as 50 (40%) are distributed in Cluster 3, while the majority of the advanced intermediate L1 Mandarin speakers rated as 50 concentrate in Cluster 1(38.11%) and Cluster 2 (42.66%). Compared with advanced intermediate speakers rated as 50, advanced speakers rated as 60 do not show as much difference based on L1 background. Most of the advanced L1 Hindi speakers rated 60 (59.38%) and advanced L1 Mandarin speakers rated 60 (63.66%) are located in Cluster 3.

Cluster 4, which includes 28 speakers in total, is a linguistic profile that invites closer examination. 12.5% of the L1 Hindi speakers rated as 50 and 9.8% of the L1 Hindi speakers rated as 60 are in cluster 4. The percentages are 9.8% for L1 Mandarin speakers rated as 50 and 9.09% for L1 Mandarin speakers rated as 60. Speakers in Cluster 4 score at the top in both fluency and vocabulary related features, whom raters might have evaluated to be advanced L2 English speakers based on fluency and vocabulary only. However, Cluster 4 includes speakers of both proficiency levels and L1 backgrounds. Other linguistic features, such as discourse structure, rhetorical patterns, and grammatical accuracy, might be influential factors worthy of exploration in future studies.

5.4. Implication for Rating and Future Pedagogy Design

This dissertation study re-emphasizes the relationship between holistic scale and analytic descriptors. Examinees who were rated the same score on a holistic scale may manifest in different linguistic profiles. During the rater training session, raters would first listen to

benchmark recordings before rating test responses. This practice will help raters gain a more comprehensive view of the holistic scale if trainers could provide responses demonstrating different linguistic characteristics.

Raters' use of holistic scales benefits from closer examination. In this dissertation, L1 mandarin speakers and L1 Hindi speakers have shown different combinations of fluency features, vocabulary features, and accentedness/effort ratings. The complementary linguistic profiles may help raters prepare and adjust during the rating process. This dissertation has provided strong evidence that raters trained to a holistic scale are sensitive to the combinations of features, very sensitive to the interplay of fluency, vocab, and accentedness/effort.

This dissertation also provides high level L2 English speakers with guidance for presentation/delivery skills. Advanced intermediate and advanced L2 English speakers in this dissertation study are not required to enroll in English courses, as their L2 English proficiency has met the "basic threshold". Mapping out different linguistic profiles, however, is still of great benefit for L2 English speakers if further progress is desired. In addition, values of fluency and vocabulary measures come to a balancing point for all the profiles. For example, speakers of higher overall L2 English proficiency were rated as less accented and cost less listener effort, which is a common phenomenon for both L1 Hindi and L1 Mandarin speakers. The purpose of this dissertation study is not asking advanced intermediate or advanced L2 English speakers to eliminate their accents by using training drills of segmental or suprasegmental elements, or spend a considerable amount of energy emulating L1 English speakers. However, more attention could be directed to the combination of fluency and vocabulary features, which might have impact on listeners' perception of accentedness. For instance, some L2 English speakers can try to speak slower, which may reduce listener effort. For other groups of L2 English speakers, increasing

delivery speed and enhancing vocabulary, may improve delivery skills and the holistic scores assigned.

5.5. Limitation of the Study

The study would greatly benefit from a more balanced database that includes the same number of advanced intermediate and advanced L2 English speakers. The number of advanced L2 English speakers, especially advanced L2 English speakers with an L1 background in Mandarin, is extremely limited. This situation is in connection with the university's admission process, where most of the admitted international graduate students scored between 80 and 100 as TOEFL total score. Compared with L1 Hindi speakers who use English as an instructional language in educational institutions, most L1 Mandarin speakers learn English as a foreign language. Different experiences and strategies in English learning have also increased the difficulty of recruiting L1 Mandarin speakers of advanced L2 English oral proficiency.

Rater training is another question to be considered for research in the future. All the raters in this study were asked to familiarize themselves with the accentedness evaluation scale and completed practice items online before they started rating. However, face-to-face community practice of rating and discussion of scale application would help raters make consensus is using rubrics. In addition, raters with professional training and experience as an ESL instructor have been exposed to a variety of English varieties and accents, which might influence their judgment on "How different the speaker's accent is from the local variety you are used to (i. e. General American English)". A rater who participated in the study described his rating experience (email communication):

> It was hard for me to answer the "what you are used to" part. My wife is a native speaker of Mandarin, so I am very used to her English. Besides her, most of my interactions are with students, so I am used to the way they speak as well. Also, I

> lived in Taiwan for over 6 years and Japan for over 6 years, so that affects what I am used to as well. Finally, the "mainstream English sample" you provided is not what I am used to very much (my parents spoke blue collar/Appalachian English when I was growing up). One of the toughest accents for me to understand was Irish when I was visiting Ireland! Yes, they are "NS" but I can understand most of my students, my wife, and my NNS colleagues a lot more easily!

The rater's experience has raised a concern about the relationship between accent and comprehensibility. Raters' familiarity with a certain accent and unfamiliarity with another may directly impact on their evaluation of "accent difference" and "listener effort required".

Also, it is impossible to complete a multi-faceted and comprehensive linguistic profile based on only the constructs of fluency, vocabulary, and accentedness. Linguistic features of other aspects will also influence the configuration of each cluster, such as grammatical accuracy and syntactic complexity. These features can be further investigated in future research in addition to fluency and vocabulary.

APPENDIX A

Python Script for the Measure of Textual Lexical Diversity

```
import os
import glob
import csv
from lexical_diversity import lex_div as ld


path = os.getcwd()
path2 = os.path.join(os.getcwd(),'text')
os.chdir(path2)
files = glob.glob('*.txt')
print(files)

# read each file and calculate each mtld
for fle in files:
    f = open(fle, 'r')
    text = f.read()
    tok = ld.tokenize(text)
    tok_num  = len(tok)
    flt = ld.flemmatize(text)

    mtld = ld.mtld(flt)
    print(mtld)


# write result into the output file
    fields = [str(fle),str(mtld),str(tok_num)]
    with open('output.csv', 'a',  newline='') as fd:
        writer = csv.writer(fd)
        writer.writerow(fields)

print(1)
```

APPENDIX B

OEPT 2 Item Summary (OEPT 2 Test Manual, p. 12)

Item no.	Item Title	Abbr	Prompt type	Expected Response
P1	Personal History 1	warm 1	Text	Talk about your country, region, or city of origin.
P2	Personal History 2	warm 2	Text	Talk about your favorite holiday in your home country.
1	Area of Study	aos	Text	Describe your area of study for an audience of people not in your field.
2	Newspaper Headline	np	Text	Given an issue concerning university education, express an opinion and build an argument to support it.
3	Compare and Contrast	cnc	Text	Based on 2 sets of given information, make a choice and explain why you made it.
4	Pros and Cons	pros	Text	Consider a TA workplace issue, decide on a course of action, and discuss the possible consequences of that action.
5	Respond to Complaint	rtc	Text	Give advice to an undergraduate concerning a course or classroom issue.
6	Bar Chart	barc	Graph	Describe and interpret numerically-based, university-related data.
7	Line Graph	lg	Graph	Describe and interpret numerically-based, university-related data.
8	Telephone Message	tel	Audio	Relay a telephone message in a voicemail to a peer.
9	Conversation	conv	Audio	Summarize a conversation between a student and prof.
10	Short lecture	sl	Audio	Summarize a lecture on a topic concerning graduate study.
11	Read Aloud 1 - Sounds	ral1	Text	Read aloud a short text containing all the major consonant and vowel sounds of English.
12	Read Aloud 2 - Text	ral2	Text	Read aloud a passage from a University policy statement containing complex, dense text.

APPENDIX C

OEPT 2 Newspaper Headline (NP) Items

Form 1: Newspaper Headline

INTRODUCTION

Next, you will see a headline from the student newspaper. As a member of the community of students and scholars at this university, you will find there are many interesting current events discussed in the local newspapers.

TASK

Your task is to express your opinion about the following newspaper headline.

The Classroom Goes Virtual: Web-based Courses Available Soon

Students will soon be able to take classes without leaving their residence halls. In the future, international students may be able to take classes without leaving their native countries.

QUESTION

Do you think that taking college courses on-line is a good way to study? Why or why not?

Form 2: **Newspaper Headline**

INTRODUCTION

Next, you will see a headline from the student newspaper. As a member of the community of students and scholars at this university, you will find there are many interesting current events discussed in the local newspapers.

TASK

Your task is to express your opinion about the following newspaper headline.

University pushes recycling with public service announcement

The University Residences is airing a public service announcement on the University Residence's cable television service, encouraging students to recycle glass, plastic, and paper products in the dorms.

QUESTION

Do you think a television announcement will have a significant effect on the amount that they recycle? Why or why not?

Form 3: Newspaper Headline

INTRODUCTION

Next, you will see a headline from the student newspaper. As a member of the community of students and scholars at this university, you will find there are many interesting current events discussed in the local newspapers.

TASK

Your task is to express your opinion about the following newspaper headline.

Undergraduate Science Class Enrolment Swells to a Record High 400

At public universities in the United States, introductory classes in science and math tend to be much larger than classes in English, History or Art.

QUESTION

Do you believe that class size affects the quality of education? Why or why not?

Form 4: Newspaper Headline

INTRODUCTION

Next, you will see a headline from the student newspaper. As a member of the community of students and scholars at this university, you will find there are many interesting current events discussed in the local newspapers.

TASK

Your task is to express your opinion about the following newspaper headline.

> ***University Takes Action Against Illegal Downloading on Campus Network***
>
> *The University has begun sending notifications to its network users who have allegedly downloaded or shared copyrighted materials (i.e. music) illegally.*

QUESTION

Do you think it is the university's responsibility to prevent students from illegally downloading music? Why or why not?

APPENDIX D

OEPT 2 Holistic Rubric

OEPT2 HOLISTIC SCALE
revised 11-8-2012

Level	General Proficiency Level	Requirements of Listener	Performance of Speaker
60	**Excellent and Consistent across items. Majority of items 60.** Minimal listener effort required to adjust to accent. Frequent displays of lexico-syntactic sophistication and fluency. Speaker is at ease and confident fulfilling task, elaborating a personalized message, using accurate English. Errors are minor and few.		
55	**More than Adequate. Mix of 55, 60, with a few 50 if any.** Little listener effort required to adjust to accent/prosody/ intonation. Consistently intelligible, comprehensible, coherent. Strong skills across items. Wide range of vocab and syntactic structures, generally sophisticated responses. Speaker may exert some noticeable effort or show minor fluency issues in elaborating clear message to fulfill task. Errors are minor.		
50	**Adequate and ready for the classroom without support. Majority of items 50, possibly some 55 or very few 45.** Acceptably small amount of listener effort required to adjust to accent/prosody/intonation. Consistently intelligible and comprehensible. Speaker may exert a little noticeable effort, but despite minor errors of grammar/vocab/stress/fluency, message is adequately coherent, with correct information, some lexico-syntactic sophistication, and displays of automaticity and fluency.		
45	**Borderline - Inconsistent – Minimally adequate for classroom with support. Mix of 45 and 50, very few, if any, 40.** Tolerable listener effort required to adjust. Consistently intelligible. Strengths & weaknesses across characteristics or items. Message is generally coherent, but may require more than a little noticeable effort for speaker to compose, or delivery may be slow. Or message may be clear and expressed fluently, but language use is somewhat simplistic.		
40	**Limited - Not ready for the classroom. Mix of 40 and 45, or a few 35, if any.** Able to address prompts and complete responses. Consistent listener effort may be necessary. Message may be simplistic/unfocussed/incomplete/ incorrect. May struggle somewhat to build sentences/argument or to articulate sounds. May be occasionally unintelligible, incomprehensible, or incoherent.		

35	**Restricted - May need more than 1 semester of support. Mix of 35 and 40.** Listening may require considerable effort. May be unintelligible or incoherent more than occasionally OR have marked deficiencies in at least 3 other areas: fluency, vocabulary, grammar/syntax, listening comprehension, articulation/pronunciation, prosody. May have difficulty completing responses.

APPENDIX E

Appendix E-1 Breakdown Fluency Measures Related to Filled Pauses

	Variable Name	Definition	Research Related	Ratio/ Normed
	Filled Pauses	Filled pauses were defined as non-contributory voiced fillers (Riggenbach, 1991), which could be: - *Non-lexical fillers* such as "uh" and "um"; - *Sound stretches* with vowel elongations of .3 seconds or longer; - *Lexical fillers* without semantic contribution such as "you know" and "I mean".	Riggenbach, 1991	
Filled Pauses	**Number of Filled Pauses**	Number of "Voiced fillers, which do not normally contribute lexical information".	Ginther, Dimova, & Yang, 2010	
	Total Filled Pause Time	The sum of time for all filled pauses. In Iwashita et al. (2008), the total filled pause time is shown as the percentage of total speaking time.	Bosker et al., 2014; Ginther, Dimova, & Yang, 2010; Iwashita et al., 2008;	
	Mean Filled Pause Length	Filled pause time divided by number of filled pauses	Ginther, Yang & Dimova, 2010	✓
	Filled Pause Ratio	Filled pause time as a decimal percent of total response time	Ginther, Dimova, & Yang, 2010	✓

	Variable	Definition	Research Related	Ratio/ Normed
Filled Pauses	**Filled Pauses per word**	The number of filled pauses, divided by number of words in the response	De Jong et al., 2012a	✓
	Filled Pauses per Second Spoken	Number of filled pauses divided by phonation time (response time excluding silent and filled pauses time)	Bosker et al., 2014	✓
	Filled Pauses per Minute	Total number of filled pauses divided by response time in minutes. 60 sec./min. times total number of filled pauses (pauses filled with uhm, mm, etc.) divided by the total time speaking in seconds.	Kormos & Denés, 2004, Kormos, 2006	✓
	Filled Pauses per T-Unit	Total number of filled pauses divided by total number of T-Units.	Lennon, 1990	✓

Appendix E-2 Breakdown Fluency Measures Related to Unfilled Pauses

Main Construct	Variable Name	Definition	Research Related	Ratio/ Normed
Unfilled pauses	**Unfilled Pauses**	Number of silent pauses of 0.5 seconds or greater	Riggenbach, 1991	
	Total Silent Pause Duration	The sum of all silent pause times. This variable is closely related with the minimum threshold of silent pause, which is determined by researchers. For examples: - A silence of 0.2 seconds or less is defined as micropause (Riggenbach, 1991); - A silence between 0.3 to 0.4 seconds is defined as hesitation (Riggenbach, 1991); - A pause is defined as a break of 1 second or longer (Skehan & Foster, 1997); - Pauses are silence intervals longer than 0.2 seconds (Kormos & Dénes, 2004) - A pause is defined as an interruption to the speech flow of more than 400 milliseconds (Skehan, Foster & Shum, 2016);	Bosker et al., 2014; Freed, Segalowtiz, & Dewey, 2004; Iwashita et al., 2008; Kormos & Denés, 2004; Riggenbach, 1991; Segalowitz et al., 2017; Skehan & Foster, 1997; Skehan & Foster, 2008; Skehan, Foster & Shum, 2016	
	Number of Silent Pauses	Number of periods of silence of at least .25 seconds. This variable is also closely related with the minimum threshold of silent pause, as the variable Total Silent Pause Duration.	Ginther, Dimova, & Yang, 2010; Goldman-Eisler, 1968;	

Main Construct	Variable	Definition	Research Related	Ratio/ Normed
Unfilled Pauses	**Mean Silent Pause Time**	Silent pause time divided by number of silent pauses.	Ginther, Dimova, &Yang, 2010 Kormos & Denés, 2004;	✓
	Silent Pause Ratio	Silent pause time as a decimal percent of total response time	Ginther, Dimova, & Yang, 2010	✓
	Mean Silent Pause Duration between AS units	Silent pause time between AS units divided by number of silent pauses between AS units	Huensch & Tracy-Ventura, 2017	✓
	Pause (silent) distribution	Ratio of silent pauses within constituent boundaries to silent pauses at boundaries The value of pausing is standardized per 100 words (Skehan & Foster, 2008)	Lennon, 1984; Möhle,1984; Riazantseva, 2001	✓

Appendix E-3 Speed Fluency Measures

Main Construct	Variable Name	Definition	Research Related	Ratio/ Normed
Speed Fluency Measures related to Speech Quantity	**Amount of Speech**	The raw frequency of total number of words or semantic units produced during the response time.	Riggenbach, 1991	
	Total Duration	The total time used to complete the speech elicitation task, including phonation time, filled pauses, and unfilled pauses.	Ginther et al., 2010; Hilton, 2009.	
	Mean Length of Run	Mean Length of Run is considered as a combination of speed fluency and breakdown fluency, and refers to the amount of speech speakers produce between pauses, reflecting "a word, a phrase, a sentence or a series of sentences depending on the task and the rate of output." (p. 40).	Grosjean, 1980	✓
	Mean Length of Syllables	The Mean Length of syllables was selected to measure speed fluency in Bosker et al. (2013) – log (Spoken time / number of syllables). A log transformation is used for normal distribution approximation. In Huensch and Tracy-Ventura (2017), Mean Syllable Duration is the inverse of Articulation Rate: phonation time (i.e. speaking time excluding pauses)/total number of syllable.	Bosker et al., 2013; Huensch & Tracy-Ventural, 2017; Towell et al., 1996;	

Main Construct	Variable Name	Definition	Research Related	Ratio/ Normed
Speed Fluency Measures related to Speech Quantity	**Articulation Rate**	This variable is also closely related with response time. In comparison to speech rate, however, the calculation of articulation rate excludes time for pausing. - Number of syllables spoken per minute excluding pause time (Raupach, 1980) - The total number of syllables produced in a given speech sample divided by the amount of time taken to produce them in seconds, which is then multiplied by 60. Unlike in the calculation of speech rate, pause time is excluded. Articulation rate is expressed as the mean number of syllables produced per minute over the amount of time spent speaking when producing the speech sample (Kormos, 2006).	Kormos, 2006 Raupach, 1980;	✓
	Pruned Number of Syllables	Number of syllables after removing self-repetitions, repairs, and other language words.	Ginther et al., (2010)	
	Pruned Syllable Duration	Syllable duration	Huensch & Tracy-Ventura (2017)	

Main Construct	Variable Name	Definition	Research Related	Ratio/ Normed
	Phonation Time	Total time actually speaking	Cucchiarini et al., 2000; Ginther et al., 2010; Towell et al., 1996	
Speed Fluency Measures related to Speech Quantity	**Syllable Run All Pauses**	Run length or mean number of syllables spoken before all pause interruptions, including both filled and unfilled pauses.	Cucchirarini et al., 2002	
	Phonation Run All Pauses	Phonation duration divided by number of filled and unfilled pauses	Segalowtiz et al., 2017	
	Phonation Run	Phonation duration divided by the number of unfilled pauses.	Segalowitz et al., 2017	

Appendix E-4 Repair Fluency Measures

	Variable Name	Explanation	Related Research	Ratio/Normed
Repair Strategies & Ratio-based Disfluencies	**Disfluencies per minute**	Number of "repetitions, restarts, and repairs" (p.152) divided by response time per minute.	Hieke, 1985;	✓
	Repetitions per T unit	Number of repeated words divided by number of T-units.	Lennon, 1990	✓
	Repetitions per 100 word	100 times the number of repetitions, divided by number of words in the response.	Yoshitomi, 1999	✓
	Self-corrections per 100 word	100 times the number of self-corrections, divided by number of words in the response.		✓
	Repetitions per second	Number of repetitions divided by response time in seconds.	De Jong et al., 2015; Huesch & Tracy-Ventura, 2017	✓
	Restarts per second	Number of restarts divided by response time in seconds.		✓
	Corrections per second spoken	Number of corrections divided by total speech time excluding pauses.	De Jong et al., 2015; Huesch & Tracy-Ventura, 2017.	✓

APPENDIX F

Lexical Diversity Measures

Reference	Construct	Measure	Calculation	Evaluation	Context
Linnarud, 1986	Lexical richness	Lexical individuality Lexical density Lexical variation Lexical sophistication	**Lexical individuality:** Words used only for one writer **Lexical density:** percentage of lexical words in total number of words **Lexical variation:** Type/token ratio **Lexical sophistication**: Words that are normally not expected at the level of instruction	Whether the concept of rare words can distinguish between native speakers outside of a classroom setting remains to be proven, or whether it could be applied to oral proficiency as well (Daller et al., 2003).	L2 K-12 /Writing
Laufer, 1994, 1995	Lexical richness	**Lexical Frequency Profile (LFP):** The LFP shows the percentage of words that learners use in different vocabulary level in their writing,	If among the 200 word types, 150 belong to the first 1000 most frequent words, 20 to the second 1000, 20 to the University Word List (Xue & Nation, 1984), and 10 not in any list, then the LFP of the composition is 75%-10%-10%-5%.	LFP is a critical index to assess lexical quality. The better a composition was, the larger the percentage of non-basic words it was expected to contain.	L2 Adult/Speaking and writing

Reference	Construct	Measure	Calculation	Evaluation	Context
Laufer & Nation, 1995	Lexical richness	**Lexical Originality (LO):** The percentage of words in a given piece of writing that are used in by one particular writer and no one else in the group, **Lexical Variation (LV**): The ratio in per cent between the different words in the text and the total number of running words—Type/token ratio **Lexical Density (LD):** The percentage of lexical words in the text (nouns, verbs, adjectives, adverbs), or content words **Lexical Sophistication (LS):** The percentage of "advanced" words in the text, which is based on users' definition.	LO = $\frac{\text{Number of tokens unique to one writer}}{\text{Number of tokens}}$ LV = $\frac{\text{Number of types} \times 100}{\text{Number of tokens}}$ LD = $\frac{\text{Number of lexical tokens} \times 100}{\text{Number of tokens}}$ LS= = $\frac{\text{Number of advanced } tokens \times 100}{\text{Number of tokens}}$	Evaluation of LV in measuring lexical richness: a. Sensitive to length and unstable for short texts b. Can be affected by differences in text length. c. Dependent on the definition of a word, or whether derivatives would be considered as new words; d. Does not distinguish vocabulary of different frequency levels. e. Unstable above a certain level of proficiency (Vermeer, 2000)	L2 Adult/Writing

Reference	Construct	Measure	Calculation	Evaluation	Context
Malvern & Richards, 1997; Yu, 2009.	Lexical diversity	TTRs calculated at different text lengths (100 tokens, 200 tokens or 300 tokens) A resulting curve demonstrating the decrease of TTR as text length increases.	**Malvern and Richard's $\mathcal{D}$**, which is a parameter determining the shape of the curve.	The effect of text length will be leveled out to some extent, but not entirely.	L2 Adult/Writing L2 Adult/Speaking
Daller et al., 2003; Guiraud, 1954	Lexical richness	The Index of Guiraud Advanced TTR Advanced Guiraud	$\text{TTR} = \frac{\text{types}}{\text{tokens}}$ $\text{G} = \frac{\text{types}}{\sqrt{\text{tokens}}}$ $\text{A}_{\text{TTR}} = \frac{\text{advanced types}}{\text{tokens}}$ $\text{A}_{\text{G}} = \frac{\text{adbanced types}}{\sqrt{\text{tokens}}}$	Evaluation of advanced TTR and advanced Guiraud: a. The square root in the denominator leads to a higher G for a longer text with the same TTR as a shorter one, which helps maintain the same TTR for a larger sample. b. Might be the most stable index for language learner data after transformation (van Hoot & Vermeer, 1988) c. Over and under adjustment may exist (Jarvis, 2002) d. Unstable above a certain level of proficiency (Vermeer, 2000)	L2 Adult/ Speaking

Reference	Construct	Measure	Calculation	Evaluation	Context
Malvern & Richards, 2007	Lexical diversity	Transformation of TTR: CTTR RTTR LogTTR	**CTTR**= $\frac{types}{\sqrt{2tokens}}$ (Carroll's lexical diversity measure/correted TTR) **RTTR** = $\frac{types}{\sqrt{tokens}}$ (Guiraud's Index/Root TTR) **LogTTR**=Log Types/Log Tokens (Herdan's Index/Bilogarithmic TTR)	None of the three measures shows advantages over others. All measures require the same sample size to be reliable.	L2 Adult/Writing
van Hout & Vermeer (2007); Vermeer (2004)	Lexical richness	**Measure of Lexical Richness (MLD)**: a lexical measure based on the relative frequency of words as they occur on a daily basis. Nine categories of vocabulary classes are established first based on geometric mean across different corpora.	**Measure of Lexical Richness (MLR)** MLR = q1 + 1/1.25*q2 +1/1.75*q3+1/2*q4+1/3*q5+ ¼*q6+1/6*q8*(4.6)+1/9*q9(13.8) q1-q9=quotient of text coverage of the transcript of speech and 'model coverage' (token coverage in Schrooten and Vermeer (1994)).	MLR can give an indication of a person's vocabulary size like an extrapolated score on a vocabulary test related to a dictionary.	L2 K-12/Speaking

Reference	Construct	Measure	Calculation	Evaluation	Context
Tidball & Treffers-Daller, 2007	Lexical richness	**Limiting Relative Diversity:** An index that distinguishes advanced words from basic words.	LRD (basic/all) = 1- $\sqrt{\mathrm{D(basic/D(all))}}$ D stands for **Malvern and Richard's** $\mathcal{D}$ (i.e. D-measurement)	It is crucial to ensure that the distinction between basic and advanced words is based on valid criteria.	L2 Adult/Speaki-ng
Henrichs & Schoonen, 2009	Lexical density Lexical diversity	**Lexical Density:** The relationship between the number of words with lexical properties as opposed to the number of words with grammatical properties. **Lexical Diversity:** The degree to which new words are introduced and used in a text.	**Lexical Density:** A percentage of the number of lexical words over all words in a text. **Lexical Diversity:** **Malvern and Richard's** $\mathcal{D}$ (i.e. D-measurement	Partial correlation was found between parental language input lexical diversity and children's vocabulary test performance. Lexical diversity is used as a predictive index of vocabulary growth.	L1 K-12 Speaking

Reference	Construct	Measure	Calculation	Evaluation	Context
Malvern & Richards, 2009	Rare word diversity	**Rare Word Diversity:** It reflects the range of sophisticated vocabulary that the speaker or writer brings to the task.	RWD (Rare Word Diversity) $= D_{All} - D_{Basic}$	RWD meets the basic requirement of producing a reasonable distribution, which differentiates among the texts being measured. RWD behaves as expected with an engineered decrease in diversity, and does indeed go negative when the deployment of rare words is less diverse than that found in the basic vocabulary of a language sample. RWD is superior to Advanced Guiraud or Advanced TTR.	L2 Adult /Writing
Skehan, 2009a, 2009b	Language sophistication	**Lambda** Higher lambda represents a wider vocabulary range	Poisson distribution is applied in explaining the construct. A text could be divided into chunks of 10 words. For each chunk, the number of difficult (threshold frequency) words is calculated. After calculating the number of chunks containing no difficult word, one difficult word or two difficult words, a statistic lambda is calculated, representing the best distribution of numbers of difficult words.	Does not correlate with $\mathcal{D}$ generated by vocd.	L2 Adult /Speaking

Reference	Construct	Measure	Calculation	Evaluation	Context
Treffers-Daller, 2009	Lexical diversity Lexical richness	**General lexical richness:** Gives a good impression of the differences in lexical diversity between text from different sources. **Diversity of individual word categories:** Analyses of particular lexical categories can help identify those contribute most to text variability.	**General lexical richness measures:** Guiraud The Index of D **Lexical diversity of nouns and verbs:** Guiraud nouns 1 (noun types/$\sqrt{\text{noun tokens}}$) Guiraud nouns 2 (noun types/$\sqrt{\text{all tokens}}$) Guiraud verbs 1 (verb types/$\sqrt{\text{verb tokens}}$) Guiraud verbs 2 (verb types/$\sqrt{\text{all tokens}}$)	More detailed representation of lexical diversity (nouns and verbs specifically) can help identify the functional category that contributes more to diversity, and discriminate better among different proficiency groups.	L2 Adult /Speaking

REFERENCES

Abercrombie, D. (1949). Teaching pronunciation. *English Language Teaching, 3*(5), 113–122.

Ågren, M., Granfeldt, J., & Schlyter, S. (2012). The growth of complexity and accuracy in L2 French: Past observations and recent applications of developmental stages. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 proficiency and performance: Complexity, accuracy and fluency in SLA* (pp. 95-119). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Anderson, A., Downs, S. D., Faucette, K., Griffin, J., King, T., & Woolstenhulme, S. (2007). How accents affect perception of intelligence, physical attractiveness, and trustworthiness of Middle-Easter-, Latin-American-, British, and Standard American-English-accented speakers. *BYU Undergraduate Journal of Psychology, 3*(1), 5–11.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*(4), 529–555.

Anderson, R. C., & Freebody, P. (1981). Vocabulary Knowledge. In J.T. Guthrie (Ed.) *Comprehension and Teaching: Research Reviews* (pp. 77-117). Newark, DE: International Reading Association.

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say thee uh you're describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(5), 914–930.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bartening, I. (2000). Gender agreement in L2 French: Pre-advanced vs advanced learners. *Studia Linguistica*, *54*(2), 225-237.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America, 114*(3), 1600–1610.

Blake, C. (2006). *The potential of text-based Internet chats for improving ESL oral fluency* (Unpublished Doctoral Dissertation). Purdue University, West Lafayette, IN.

Bourdieu, P., & Thompson, J. B. (1991). *Language and symbolic power*. Cambridge, MA: Harvard University Press.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159-175.

Bosker, H. R., Quené, H., Sanders, T. J. M., & De Jong, N. H. (2014). The perception of fluency in native and non-native speech. *Language Learning*, *64*(3), 579-614.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohens (Eds.), *Interface between second language acquisition and language testing* (pp. 112-140), Cambridge, UK: Cambridge University Press.

Bulté, B., & Housen, A. (2012). Defining and opertationalising complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 proficiency and performance: Complexity, accuracy and fluency in SLA* (pp. 47-69). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bygate, M. (2001). Effects of task repetition on the structure and control of language. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, Teaching and Testing* (pp. 23-48). London, UK: Routledge.

Callies, M., Diez-Bedmar, M. B., & Zaytseva, E. (2014). Using learner corpora for testing and assessing L2 proficiency. In P. Leclerq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp, 71-90). Bristol, UK: Multilingual Matters.

Canale, S., & Swain, M. (1980). Theoretical bases of communicative language approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Cao, H. (2014). *Disentangling Fluency, Comprehensibility and Coherence: Towards a Better Understanding of Oral English Profiles*. (Unpublished Doctoral Dissertation). Purdue University, West Lafayette, IN.

Carroll, J. B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535-544.

Clahsen, H. (1985). Profiling second language development: A procedure for assessing second language proficiency. In K. Hiltenstam & M. Piennemann (Eds.), *Modelling and assessing second language acquisition* (pp. 283-331). Clevedon, UK: Multilingual Matters.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658.

Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: the temporal delay hypothesis. *Plos One*, *6*: e19792. Available from: https://doi.org/10.1371/journal.pone.0019792

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Coxhead, A. (2016). Reflecting on Coxhead (2000): "A new academic word list". *TESOL Quarterly*, *50*(1), 181-185.

Crossley, S. A., & McNamara, D. N. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*(2), 119-135.

Crossley, S. A., Salsbury, T., & McNamara, D. N. (2009). Measuring L2 lexical growth using hypernymic relations. *Language Learning*, *59*(2), 307-334.

Crossley, S. A., & Salsbury, T. (2010). Using lexical indices to predict produced and not produced words in second language learners. *The Mental Lexicon*, *5*(1), 115-147.

Crossley, S., Salsbury, T., McNamara, D., & Jarvis, S. (2011a). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*(1), 182-193.

Crossley, S., Salsbury, T., Mcnamara, D., & Jarvis, S. (2011b). Predicting language proficiency in language learner text using computational indices. *Language Testing*, *28*(4), 561-580.

Crossley, S. A., Salsbury, T., & McNamara, D. N. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243-263.

Cucchiarini, C., Strik, H., & Boves, L. W. J. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of Acoustic Society of America, 107*(2), 989-999.

Cucchiarini, C., Strik, H., & Boves, L.W. J. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*(6), 2862-2873.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. London, UK: Blackwell.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals, *Applied Linguistics*, *24*(2), 197-222.

De Bot, K. (1992). A bilingual production model: Levelt's "speaking" adapted. *Applied Linguistics*, *13*(1), 1-24.

Deese, J. (1980). Pauses, prosody and the demand in language production. In H. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 3-10). The Hague: Mouton.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, R. (2012a). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 proficiency and performance: Complexity, accuracy and fluency in SLA* (pp. 121-142). Amsterdam/Philadelphia: John Benjamins Publishing Company.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. (2015). Second language fluency: Speaking style or fluency? Correcting second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223-243.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1–16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476–490.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam/Philadelphia: John Benjamins.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgements on different tasks. *Language Learning*, *54*(4), 655–679.

Divjak, D. T., & Gries, S. (2006). Ways of trying in Russian: Clustering. *Corpus linguistics and linguistic theory*, *2*(1), 23-60.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition, 20*(3), 349-385.

Duran, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220-242.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139-155.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. *Cluster Analysis*, Chichester, UK: Wiley.

Faerch, K., Haastrup, K., & Phillipson, R. (1984). *Learner language and language learning*. Copenhagen, Denmark: Multilingual Matters.

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler, & W. S-Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior* (pp. 85-101). New York, NY: Academic Press.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America, 76*(3), 692–707.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology Human Perception and Performance, 32*(5), 1276–1293.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency and lexical diversity. *Language Learning*, *59*(4), 866-896.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied linguistics*, *21*(3), 354-375.

Fraser, C., & Kelly, B. F. (2012). Listening between the lines: Social assumptions around foreign accents. *Australian Review of Applied Linguistics, 35*(1), 74–93.

Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In. B. Freed (Ed.) *Second language acquisition in a study abroad context* (pp. 123-148). Philadelphia, PA: John Benjamins.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243-265). Ann Arbor, MI: University of Michigan Press.

Freed, B. F., Segalowitz, N., & Dewey, D. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*. *26*(2), 273-301.

Friginal, E., Li, M., & Weigle, S. C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays, *Journal of Second Language Writing*, *23*(1), 1-16.

Fuertes, J. N., Potere, J. C., & Ramirez, K. Y. (2002). Effects of speech accents on interpersonal evaluations: Implications for counseling practice and research. *Cultural Diversity & Ethnic Minority Psychology, 8*(4)*,* 346–356.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65–89.

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science, 28*(6), 703–712.

Ginther, A., Dimova, S., & Yang, R. Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implication for automated scoring. *Language Testing*, *27*(3), 377-399.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*(2), 271-295.

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*(2), 96-102.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London, UK: Academic Press.

Grandfeldt, J. & Ågren, M. (2014). SLA developmental stages and teachiers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing*, *31*(3), 285-305.

Gries, S. T. (2010). Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon*, *5*(3), 323-346.

Grosjean, F. (1980). Temporal variables within and between languages. In H. W. Dechert & M. Raupach (Eds.), *Towards a cross-linguistic assessment of speech production* (pp. 39-53). Frankfurt, Germany: Peter Lang.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: Essai de méthodologie*. Paris, France: Presses Universitaires de France.

Guiraud, P. (1960). *Problèmes et Méthodes de la Statistique Linguistique*. Dordrecht, the Netherlands: D. Reidel.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201–223.

Harding, L. (2017). Validity in pronunciation assessment. In O. Kang & A. Ginther (Eds.). *Assessment in second language pronunciation* (pp. 30–48). Oxfordshire, UK: Routledge.

Harmer, J. (1991). *The practice of English language teaching*. New York, NY: Longman.

Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech*, *14*(3), 277-288.

Hayes-Harb, R. (2014). Acoustic-phonetic parameters in the perception of accent. In A. Moyer & J. Levis (eds.), *Social dynamics of second language accent* (pp. 31-51). London, UK: Blackwell.

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam/Philadelphia: John Benjamins.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies of Second Language Acquisition*, *21*(2), 303-317.

Henrichs, L. & Schoonen, R. (2009). Lexical features on parental academic language input: The effect on vocabulary growth in monolingual Dutch children. In Richards, B., Daller, M. H., Malvern, D. D., Meara, P., Milton, J. and Treffers-Daller, J. (Eds.), *Vocabulary studies in first and second language acquisition* (pp. 1-22). Basingstoke, UK: Palgrave Macmillan.

Herdan, G. (1960). *Quantitative linguistics*. London, UK: Butterworth.

Hieke, A. E. (1985). A componential approach to oral fluency evaluation. *Modern Language Journal*, *69*(2), 135-142.

Hieke, A. E., Kowal, S., & O'Connell, D. C., (1983), The trouble with "articulatory" pauses, *Language and Speech*, *26*(3), 203-214.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, *36*(2), 153-166.

Hilton, H. (2009). Annotation and analyses of temporal aspects of spoken fluency, *CALICO Journal*. *26*(3), 644-661.

Hosoda, M., Stone-Romero, E. F., & Walter, J. N. (2007). Listeners' cognitive and affective reactions to English speakers with standard American English and Asian accents. *Perceptual and Motor Skills, 104*(1), 307–26

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy, and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 proficiency and performance: Complexity, accuracy and fluency in SLA* (pp. 1-20). Amsterdam/Philadelphia: John Benjamins.

Huensch, A., & Tracy-Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: a multidimensional investigation, *Modern Language Journal*, *101*(2), 275-293.

Hymes, D. (1972). On communicative competence. In J. B. Pride & A. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth, UK: Penguin.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*(3), 475–505.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24-49.

Jarvis, S. (2002). Short text, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57-84.

Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392.

Jarvis, S. (2012). Lexical challenges in the intersection of applied linguistics and ANLP. In C. Boonthum-Denecke, P. M. McCarthy, & T. Lamkn (Eds.), *Cross disciplinary advances in applied natural language processing: Issues and approaches* (pp. 50-72). Hershey, PA: IGI Global.

Jarvis, S. (2013a). Capturing the diversity in lexical diversity. *Language Learning*, *63*, Suppl.1, 87-106.

Jarvis, S. (2013b). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13-44). Amsterdam/Philadelphia: John Benjamins.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377-403.

Johnson, W. (1939). *Language and speech hygiene: An application of general semantics*. Ann Arbor, MI: Edward Brothers.

Johnson, W. (1944). Studies in language behavior: I. A program of Research. *Psychological Monographs*, *56*(2), 1-15.

Juan-Garau, M. (2014). Oral accuracy growth after formal instruction and study abroad: Onset level, contact factors and long-term effects. In C. Pérez-Vidal (Ed.), *Language acquisition in study abroad and formal instruction contexts* (pp. 87-110). Philadelphia/Amsterdam: John Benjamins.

Juan-Garau, M. (2018). Exploring oral L2 fluency development during a three-month stay abroad through a dialogic task. In C. Sanz & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 193-208). New York, NY: Routledge.

Kalin, R., & Rayko, K. (1978). Discrimination in evaluative judgements against foreign-accented job candidates. *Psychological Reports, 43*, 1203–1209.

Kang, O. (2008). Ratings of L2 oral performances in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 181-205.

Kang, O., & Rubin, D. (2014). Listener expectations, reverse linguistic stereotyping, and individual background factors in social judgments and oral performance assessment. In A. Moyer & J. Levis (Eds.). *Social dynamics in second language accent* (pp. 239-253). Boston/Berlin: Mouton de Gruyter.

Keβler, J. U., & Liebner, M. (2011). Diagnosing L2 development: Rapid profile. In M. Pienemann & J. U. Keβler (Eds.), *Studying processability theory: An introductory Textbook* (pp. 133-148). Amsterdam/Philadelphia: John Benjamins.

Kinzler, K. D., & DeJesus, J. M. (2013). Northern = smart and Southern = nice: The development of accent attitudes in the United States. *The Quarterly Journal of Experimental Psychology, 66*(6), 1146–1158

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research, 4*(5), 900–913.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). Ann Arbor, MI: University of Michigan Press.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners, *System, 32*(2), 146-164.

Kowal, S., O'Connell, D., & Sabin, E. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research*, *4*(3), 195-207.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal*, *70*(4), 366–372.

Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 proficiency and performance: Complexity, accuracy and fluency in SLA* (pp. 1-20). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kuznetsova, J. (2015). *Linguistic profiles: Going from form to meaning via statistics*. The Hague: Morton.

Labov, W. (2006). *The social stratification of English in New York City* (2nd ed.). Cambridge, UK: Cambridge University Press.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. Ann Arbor, MI: University of Michigan Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. New York, NY: McGraw Hill.

Lado, R. (1964). *Language teaching: A scientific approach*. New York, NY: McGraw-Hill.

Lambert W. E., Hodgson, R., Gardner, R., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology, 60*(1), 44–51.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied linguistics*, *27*(4), 590-619.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21-33.

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. In J. Eubank, L. Selinker & M. Sharwood Smith (Eds.), *The current state of interlanguage: Studies in honor of William E. Rutherford* (pp. 265-272). Philadelphia/Amsterdam: John Benjamins.

Laufer, B.. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33-51.

Lennon, P. (1990). Investigating fluency in EFL: A Quantitative Approach. *Language Learning*, *40*(3), 387-417.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Ann Arbor, MI: University of Michigan.

Levelt, W. J. M. (1989). *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. M. (1999a). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp.83-122). Oxford, UK: Oxford University Press.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe nonnative speakers? The influence of on credibility. *Journal of Experimental Social Psychology, 46*(6), 1093–1096.

Levin, H., & Silverman, I. (1965). Hesitation phenomena in children's speech. *Language and Speech*, *8*(2), 67-85.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369-377.

Linnarud, M. (1986*). Lexis in composition: A performance analysis of Swedish learners' written English*. Malmo, Sweden: Liber Forlag (CWK Gleerup).

Livingston, B. A., Schilpzand, P., & Erez, A. (2017). Not what you expected to hear: Accented messages and their effect on choice. *Journal of Management, 43*(3), 804–833.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Mackey, A., & Gass, S. (2012). *Research methods in second language acquisition: A practical guide*. Oxford, UK: Wiley-Blackwell.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*(1), 19-44.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*(4), 539–556.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon, UK: Multilingual Matters.

Malvern, D. D., & Richards, B. J. (2002). Investigating accommodation to language proficiency interview using a new measure of lexical diversity. *Language Testing*, *19*(1), 85-104.

Malvern, D. D., & Richards, B. J. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459-488.

Malvern, D. D., & Richards, B. J. (2009). A new method of measuring rare word diversity: The example of L2 learners of French. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, P., Milton, J., & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition* (pp. 164-178). Basingstoke, UK: Palgrave Macmillan.

Malvern, D. D., Richards, B. J., Ghipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459-488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392.

Meara, P. (1996a). The dimensions of lexical competence. In G. Brown, K. Malmkjoer, & Williams, J. (Eds.). *Performance and competence in second language acquisition* (pp. 33-53). Cambridge, UK: Cambridge University Press.

Meara, P. (2005). Designing vocabulary tests for English, Spanish, and other languages. In C. Butler, S. Christopher, M. A. Gomez-Gonzales, & S. M. Doval-Suarez (Eds.), *The dynamics of language use: Functional and contrastive perspectives* (pp. 271-285). Amsterdam/Philadelphia: John Benjamins.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy* (pp. 84-102). Cambridge, UK: Cambridge University Press.

Möhle, D. (1984). A comparison of the second language speech production of different native speakers. In Dechert, H. W., Möhle, D., & Raupach, M. (Eds.), *Second language productions* (pp. 26–49). Tubingen, Germany: Gunter Narr.

Morley, J. (1994). *Pronunciation pedagogy and theory: New views, new directions*. Alexandra, VA: TESOL.

Moyer, A. (2013). *Foreign accents: The phenomenon of non-native speech*. Cambridge, UK: Cambridge University Press.

Moyer, A., & Levis, J. (Eds.). (2014). *Social dynamics of second language accent*. London, UK: Blackwell.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*(3), 289–306.

Munro, M. J., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, *44*(3), 316–327.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York, NY: Newbury House Publishers.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nelson, L. R., Signorella, M. L., & Botti, K. G. (2016). Accent, gender, and perceived competence. *Hispanic Journal of Behavioral Sciences*, *38*(2), 166–185.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics*, *30*(4), 555-578.

Nortier, J. (1989). *Dutch and Moroccan-Arabic in contact: code-switching among Moroccans in Netherlands* (Unpublished Doctoral Dissertation). University of Amsterdam, the Netherlands.

Ockey, G. J. (2018). Reliability and sources of score variance in the strength of an accent scale. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 Listening: Moving towards authenticity*. (pp. 83-95), Amsterdam/Philadelphia: John Benjamins.

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, *37*(5), 693-715.

Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of strength of accentedness on an L2 interactive lecture listening comprehension test. *International Journal of Listening*, *30*(1-2), 84-98.

Ockey, G. J., & Wagner, E. (2018). An overview of using different types of speech varieties. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 Listening: Moving towards authenticity*. (pp. 67-81), Amsterdam/Philadelphia: John Benjamins.

O'Connell, D. C., & Kowal, S. (1980). Prospectus for a science in pausology. In H. Dechert & M. Raupach (Eds.). *Temporal variables in speech: Studies in Honour of Frieda Goldman-Eisler* (pp. 3-10). The Hague: Mouton Publishers.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590-601.

Park. S. (2016). *Measuring fluency: Temporal variables and pausing patterns in L2 English speech*. (Unpublished Doctoral Dissertation). Purdue University, West Lafayette, IN.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191–226). London, UK: Longman.

Pennington, M., & Richards, J. (1986). Pronunciation revisited. *TESOL Quarterly, 20*, 207–225.

Pérez-Vidal, C., & Juan-Garau, M. (2011). The effect of context and input conditions on oral and written development: A Study Abroad perspective. *International Review of Applied Linguistics in Language Teaching*, *49*(2), 157-185.

Phakiti, A. (2018a). Exploratory factor analysis. In A. Phakiti, De Costa, P., Plonsky, L. & Starfield, S. (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 423-457). London, UK: Palgrave Macmillan.

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam/Philadelphia: John Benjamins.

Pienemann, M. (Ed.) (2005). *Cross-linguistic aspects of processability theory*. Amsterdam/Philadelphia: John Benjamins.

Pienemann, M. & Keβler, J. U. (Eds.) (2011). *Studying processability theory: An introductory textbook*. Amsterdam/Philadelphia: John Benjamins.

Raupach, M. (1980). Temporal variables in first and second language production. In H. Dechert & M. Raupach (Eds.). *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 263-270). The Hague: Mouton Publishers.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge, *Language Testing*, *10*(3), 355-371.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209-227), Amsterdam/Philadelphia: John Benjamins.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*(3), 437-470.

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828-848.

Riazantseva, A. (2001). Second language proficiency and pausing A study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*(4), 497-526.

Richards, J. C. (1976). The role of vocabulary teaching, *TESOL Quarterly*, *10*(1), 77-89.

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423-441.

Robinson, P. (2001c). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics*, *22*(1), 27-57.

Rossiter, M. J. (2009). Perception of L2 fluency by native and non-native speakers. *The Canadian Modern Language Review/La Revue Canadiennes des Langues Vivantes*, *65*(3), 395-412.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English speaking teaching assistants. *Research in Higher Education*, *33*(4), 511-531.

Rubin, D. L. (2012). The power of prejudice in accent perception: Reverse linguistic stereotyping and its impact on listener judgments and decisions. In J. Levis & K. LeVelle (Eds.). *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference*, Sept. 2011. (pp. 11-17). Ames, IA: Iowa State University.

Rubin, D., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic of undergraduate perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations*, *14*(3), 337-353.

Ryslewicz, J. (2008). Cognitive profiles of (un)successful FL learners: A cluster analytical study. *Modern Language Journal*, *92*(1), 87-99.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstok, UK Palgrave Macmillan.

Scovel, T. (1988). *A time to speak: A psycholinguistic investigation into the critical period for human speech*. New York, NY: Harper and Row.

Segalowitz, N. (2001). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 200-219). Ann Arbor, MI: University of Michigan.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Segalowtiz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics*, *54*(2), 79-95.

Segalowitz, N., &. Freed, B. F. (2004). Context, contact, and cognition of oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(2), 173-199.

Segalowitz, N., French, L., & Guay, J. (2017). What features best characterize adult second language utterance fluency and what do they reveal about fluency gains in short-term immersion? [Special Issue]. *The Canadian Journal of Applied Linguistics*, *20*(2), 90-116.

Segalowitz, N., Watson, V. & Segalowitz, S. (1995). Vocabulary skill: Single-case assessment of automaticity of word recognition in a timed lexical decision task. *Second Language Research*, *11*(2), 121-136.

Sichel, H. S. (1971). On a family of discrete distributions particularly suited to present long-tailed frequency data. In N. F. Laubscher (Ed.), *Proceedings of the Third Symposium on Mathematical Statistics*. SACSIR, Pretoria, 51-97.

Sichel, H. S. (1975). On a distributive law for word frequencies. *Journal of the American Statistical Association*, *70*(351), 542–547.

Siegman, A. W. (1979). Cognition and hesitation in speech. In A. W, Siegman & S. Feidsten (Eds.), *Of speech and time: Temporal speech patterns in interpersonal contexts* (pp. 151-178). Hillsdale, NJ: Lawrence Erlbaum Associates.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 688.

Skehan, P. (1996). Second language acquisition and task-based instruction. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 17-30). Oxford, UK: Heinemann.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

Skehan, P. (2003). Task-based instruction. *Language Teaching*. *36*(1), 1-14.

Skehan, P. (2009a). Lexical performance by native and non-native speakers on language learning tasks. In B. Richards, H. M. Daller, D. Malvern, P. Meara, J. Milton & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: Interface between theory and application* (pp.107-124). Basingstoke, UK: Palgrave McMillan.

Skehan, P. (2009b). Modeling second language performance: Investigating complexity, accuracy, fluency and lexis. *Applied Linguistics*. *30*(4), 510-532.

Skehan, P., & Foster, P. (1996). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*, 185-211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*, 93-120.

Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: a meta-analysis of the Ealing research, in S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds). *Complexity, accuracy, and fluency in second language use, learning, and teaching* (pp. 207-226). Brussels, Belgium: University of Brussels Press.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency, *International Language Review of Applied Linguistics in Language Teaching*, *54*(2), 97-111.

Smith, R. A., Strom, R. E., & Muthuswamy, N. (2005). Undergraduates' rating of domestic and international teaching assistants: Timing and data collection and communication intervention. *Journal of Intercultural Communication and Research*, *34*(1), 3-21.

Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advanced quantitative methods in second language research* (pp. 243-274). New York, NY: Routledge.

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogical intervention. *TESOL Quarterly*, *60*(2), 447-471.

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. Language Learning, 61, Suppl. 1, 37-72.

Tavakoli, P., & Skehan, P. (2005), Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-276). Amsterdam/Philadelphia: John Benjamins.

Tidball, F., & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech. . In H. M. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 133-149). Cambridge, UK: Cambridge University Press.

Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 proficiency and performance: Complexity, accuracy and fluency in SLA* (pp. 221-245). Amsterdam/ Philadelphia: John Benjamins.

Towell, R., Hawkins, R., & Bazergut, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84-115.

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. *International Review of Applied Linguistics*, *40*(2), 117-150.

Treffers-Daller, J. (2009). Language dominance and lexical diversity: How bilinguals and L2 learners differ in their knowledge and use of French lexical and functional items. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.) *Vocabulary studies in first and second language acquisition* (pp. 74-90). Basingstoke, UK: Palgrave Macmillan.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies of Second Language Acquisition, 28*(1), 1–30.

Valls-Ferrer, M., & Mora, J. (2014). L2 fluency development in formal instruction and study abroad: The role of initial fluency level and language contact. In C. Pérez-Vidal (Ed.), *Language acquisition in study abroad and formal instruction contexts* (pp. 111-136). Philadelphia/Amsterdam: John Benjamins.

van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In H. M. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). Cambridge, UK: Cambridge University Press.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65-83.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 173-189), Amsterdam/Philadelphia: John Benjamins.

Weatherhead, D., & White, K. S. (2018). And then I saw her race: Race-based expectations affect infants' word processing. *Cognition*, *177*, 87–97.

West, M. 1953. *A General Service List of English Words*. London, UK: Longman, Green and Co.

Winters, S., & O'Brien, M. G. (2012). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication, 55*(3), 486–507

Wolfe-Quitero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawai'i Press.

Xue, G. & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215-229.

Yan, X., & Ginther, A. (2017). Listeners and raters: Similarities and differences in the evaluation of accented speech. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 67–88). Oxfordshire, UK: Routledge.

Yan, X., Kim, H. R., & Kim, J. Y. (2018). Complexity, accuracy and fluency features of speaking performances on Aptis across different CEFR levels. In V. Berry (Ed.), *ARAGs research reports online* (Report #AR-A/2018/1).

Young, R. F. (1999). Sociolinguistic approaches to SLA. *Annual Review of Applied Linguistics*, *19*, 105–132.

Young, R. F. (2008). *Language and interaction: An advanced source book*. New York, NY: Routledge.

Young, R. F. (2011). Interactional competence in language learning, teaching and testing. In E, Hinkel (Ed.), *Handbook of research in second language teaching and learning volume II* (pp. 426–443). New York, NY: Routledge.

You, Y. (2014). *Relationships between Lexical Proficiency and L2 Oral Proficiency*. (Unpublished Doctoral Dissertation). Purdue University, West Lafayette, IN.

Yoshitomi, A. (1999). On the loss of English as a second language y Japanese returned children. In L. Hansen (Ed.), *Second language attrition in Japanese contexts* (p. 80-112). New York, NY: Oxford University Press.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*, 236-259.

VITA

Jie Gao

gao339@purdue.edu

Ph.D Candidate, Second Language Studies
Department of English
Purdue University
West Lafayette, IN 47906

Education

Doctor of Philosophy in English **May 2020**
Purdue University, West Lafayette, Indiana

Master of Arts in Applied Linguistics **June 2015**
Tsinghua University, Beijing, P. R. C.

Bachelor of Arts in English (with honor and high distinction) **June 2012**
Minor in French (with high distinction)
Shandong University, Jinan, P. R. C.

Publications

Peer-reviewed Article and Book Chapter

Gao, J. (in preparation). Linguistic profiles of advanced intermediate and advanced L2 English speakers: Fluency, vocabulary, and accentedness. *Language Assessment Quarterly*.

Gao, J. & Banat, H. (in preparation). Embedded assessment in ESL program administration. *College English*.

Gao, J., Picoral, A., Macdonald, L., & Staples, S. (under review). Citation practices of L2 writers in First-Year Writing courses: Form, function, and connection with pedagogical materials. *Writing & Pedagogy*.

Gao, J., & Ginther, A. (2020). L2 speaking: Theory and practice. In Jeon, E. H. & In'nami, Y. (Eds.), *Understanding L2 Proficiency: Theoretical and Meta-Analytic Investigations*. Amsterdam/Philadelphia: John Benjamins.

Textbook

Yang, Y. & **Gao, J.** (2016), *Big Words Made Easy—a College English Vocabulary Textbook*, Beijing: Higher Education Press.

Technical Manuals

Crouch, D. C. & **Gao, j.** (2019). *The Assessment of College English/International (ACE-In) Test Specification Revision for Elicited Imitation*. West Lafayette, IN: Purdue Language and Cultural Exchange Program.

Gao, J. (2017). Reliability and Comparability of OEPT Scores. In Kauper, N., Ginther, A. & Yan, X. (Eds.), The Oral English Proficiency Test (OEPT) Technical Manual (pp. 22-29). West Lafayette, iN: Purdue University Oral English Proficiency Program.

Conference Presentations

Gao. J., Yan, Y., & Dilger, B. (2020, March). *Linking a corpus & repository for research, teaching, and professional development*. To be presented at the Digital Futures Symposium, West Lafayette, IN, U. S. A.

Gao. J., Crouch, D., & Cheng, L. (2019, October). *Concept-mapping for Guiding Rater Training in an ESL Elicited Imitation Assessment Task*. Presented at the 21st Conference of Midwest Association of Language Testers, Bloomington, IN, U. S. A.

Shin, J., Staples, S., **Gao, J.**, Swatek, A., & Picoral, A. (2018, October). *The Corpus and Repository of Writing: An Interactive Interface for Multiple Users*, Workshop presented at Writing Research without Walls Symposium, West Lafayette, IN, U. S. A.

Gao, J. (2018, September). *Linguistic Profile Analysis of Higher-Level L2 English Speakers—Chinese and Indian Examinees of a Local Test*. Presented at the 20th Conference of Midwest Association of Language Testers, Madison, WI, U. S. A.

Staples, S., Picoral, A., Shin, J., Velazquez, A., & **Gao, J.** (2018, July). *Exploring Variation and Intertextuality in L2 Undergraduate Writing in English: Using the Corpus and Repository of Writing Online Platform for Research and Writing*. Workshop presented at the 13th Teaching and Language Corpora (TALC) Conference, Cambridge, U. K.

Gao, J. (2018, July). *Processing Survey Data by Using Rasch—a Synthetic Literature Review*. Presented at Pacific-Rim Objective Measurement Symposium (PROMS), Shanghai, P.R.C.

Gao, J., Macdonald, L., Wang, Z., Picoral, A., & Staples, S. (2018, March). *Citation Practices of L2 Writers in First-year Writing Courses: Form, Function, and Connection with Pedagogical Materials*. Presented at American Association of Applied Linguistics 2018, Chicago, IL, U. S. A.

Gao, J., Banat, H., & Bushner, A. (2017, June). *Users of a Web-based Writing Repository: a Needs Analysis Survey*, Panel presented at Computers and Writing 2017. Findlay, OH, U. S. A.

Gao, J., Macdonald, L., Wang, Z., Picoral, A., & Staples, S. (2017, July). *Variability in citation practices of developing L2 writers in first-year writing courses*. Poster presented at Corpus Linguistics, Birmingham, U. K.

Banat, H., **Gao, J.**, Lan, G., Staples, S., & Dilger, B. (2017, March). *Developing a Corpus of L2 Writing and Repository of Pedagogical Artifacts: Methodology, Usability and Research*. Poster presented at American Association of Applied Linguistics 2017, Portland, OR, U. S. A.

Gao, J., Macdonald, L., and Craig, S. (2017, March). *Building a Better Team: Interdisciplinary Research and Collaboration in the Crow Project*. Panel presented at Purdue Languages & Cultures Conference 2017, West Lafayette, IN, U. S. A.

Gao, J. (2016, October). *A Comparison of Two Automatic Evaluation Tools in China*. Paper presented at the 18th Conference of Midwest Association of Language Testers, West Lafayette, IN, U. S. A.

Shin, J., Ge, L., **Gao, J.**, Partridge, R. S. & Staples, S. (2016, September). *The Effectiveness of "Soft" DDL on Reporting Verbs in L2 writing: A Corpus-based Study*. Presented at 2016 Conference of American Association of Corpus Linguistics and Technology for Second Language Learning, Ames, IA, U. S. A.

Gao, J. & Craig, S. (2016, March). *The Design and Research Potential of Crow (Corpus and Repository of Writing) for Language Research and Teaching*. Paper presented at Purdue Languages & Cultures Conference 2016, West Lafayette, IN, U. S. A.

Awards and Scholarships

College of Liberal Arts Scholarship, Purdue University	Fall 2019-Spring 2020
College of Liberal Arts Scholarship, Purdue University	Fall 2017-Spring 2018
Emerging Scholars Award, Department of English, Purdue University	Spring 2017
The Quintilian Award for Top Ten Percent Instructor Evaluations, Department of English, Purdue University	Summer 2018 Spring 2016
The Quintilian Award in Recognition of an Outstanding Commitment to Continuing Development as a Teacher Scholar, Department of English, Purdue University	Fall 2015

Grants and Funding

Gao, Jie. (2019-2020, Graduate Student Researcher). **American Council of Learned Societies Digital Extension Grant, for "Expanding the Corpus & Repository of Writing: An Archive of Multilingual Writing in English"** (PI: Dr. Bradley Dilger, Purdue University and Dr. Shelley Staples, University of Arizona, $149,663)	2019-2020
Gao, Jie. (2017-2019, Graduate Student Researcher). **Humanities without Wall (Mellon Funded Project) Changing Climate Grant, for "Crow: the Corpus & Repository of Writing** " (PI: Dr. Bradley Dilger, Purdue University, Dr. Bill Hart-Davidson, Michigan State University, and Dr. Shelley Staples, University of Arizona, $142,000)	2017-2019
Gao, Jie. (2019 June). **Purdue Graduate School Summer Research Grant for Dissertation *"Linguistic Profiles of High Proficiency Mandarin and Hindi Second Language Speakers of English".*** $3, 333	Summer, 2019

Gao, Jie. (2018 July). **College of Liberal Arts PROMISE Grant.** Travel to Pacific-Rim Objective Measurement Symposium (PROMS), Shanghai, P. R. C. $1, 500. Summer, 2018

Gao, Jie. (2018 March). **Purdue Graduate Student Government Travel Grant**. Travel to American Applied Linguistics, Chicago, IL, U. S. A. $250. Spring, 2018

Gao, Jie. (2017). **Introductory Composition at Purdue Travel Grant.** Travel to Computer and Writing Conference, Findlay, OH, U. S. A. $200. Summer, 2017

Gao, Jie. (2017 July). **College of Liberal Arts PROMISE Grant.** Travel to Corpus Linguistics at Birmingham, U.K. $1, 500. Summer, 2017

Gao, Jie. (2017). **College of Liberal Arts PROMISE grant.** Travel to American Association of Applied Linguistics, Portland, OR, U. S. A. $750. Spring, 2017