

CHARACTERIZING ORAL PROFICIENCY AND LANGUAGE USE OF  
LONG-TIME LEARNERS OF CHINESE AS AN ADDITIONAL LANGUAGE  
USING COMPUTER TECHNOLOGY

by

ELLIOTT YATES

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Asian Studies)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2012

© Elliott Yates, 2012

## **Abstract**

This thesis is an investigation into characterizing the oral language proficiency of longtime Anglo-Canadian learners of Chinese as an Additional Language (CAL) using computer technology. Semi-structured, informal, low-stakes interviews with five CAL learners were transcribed, segmented using freely available Chinese text parsing software, and analyzed using the methodologies from the field of complexity, accuracy, and fluency (CAF) studies in second language acquisition. This thesis describes operationalization of these CAF indices in the CAL context, compares the results to participant self-assessments using the Common European Framework of Reference for Languages (CEFR), and makes recommendations for adoption of CAF measurement standards and future work in characterizing language proficiency in the realms of applied linguistics research and language pedagogy in the CAL context.

## Preface

Parts of Chapters 3 and 4 were developed in collaboration with the research group of the Centre for Research in Chinese Language and Literacy Education (CRCLLE) at the University of British Columbia (UBC) led by Dr. Patricia Duff, comprising Dr. Duff, Tim Anderson, Roma Ilnyckyj, Ella Lester (VanGaya), Rachel Wang, and myself (Elliott Yates). The group conceived the idea of the interviews, for which Rachel Wang, then a full-time coordinator and research associate in the research centre, was the interviewer and primary transcriber; Rachel received assistance from Lorita Chiu in transcribing the 2010 interview data.

The research team examined data outside the scope of this thesis, but as part of our work we discussed possible ways to assess proficiency in a relatively objective fashion, instead of using standardized tests; we decided on Common European Framework of Reference for Languages (CEFR) self-assessments and various techniques examining lexical and syntactic features as well as speech fluency. It was my task to operationalize these techniques; thus the processing of the transcripts and all subsequent analysis was my work, with the exception of the fluency measurements, which were calculated by Rachel with some assistance from me and Dr. Duff (the fluency indices and analysis in this thesis are mine, however). Dr. Duanduan Li of the UBC Department of Asian Studies provided valuable ideas and feedback during the development of the linguistic analysis, and her help was invaluable to ensure the validity of the analytical techniques chosen.

Parts of this research will appear in a published monograph co-authored with the research team listed above, tentatively entitled “Learning Chinese: Linguistic, sociocultural, and narrative perspectives” (Duff, P., Anderson, T., Ilnyckyj, R., Lester (VanGaya) E., Wang, R., & Yates, E., forthcoming).

This study was approved on October 11, 2011, by The University of British Columbia Office of Research Services Behavioural Research Ethics Board, with certificate number H11-01592.

## Table of Contents

Abstract.....	ii
Preface .....	iii
Table of Contents.....	iv
List of Tables .....	vi
List of Figures .....	vii
Acknowledgements.....	viii
Chapter 1. Introduction .....	1
Chapter 2. Research in proficiency assessment through complexity, accuracy, and fluency measures.....	4
2.1 Introduction .....	4
2.2 Measuring proficiency .....	4
2.2.1 Standardized assessments in the CAL context.....	6
2.2.2 Specialized assessment methods.....	9
2.2.3 Issues with standardized and specialized assessment methods .....	10
2.2.4 Characterizing proficiency: Complexity, accuracy, and fluency (CAF) studies.....	11
2.3 CAF in the CAL context .....	18
2.3.1 CAF measures in the CAL context: Complexity .....	23
2.3.2 CAF measures in the CAL context: Accuracy .....	28
2.3.3 CAF measures in the CAL context: Fluency .....	31
2.4 Gaps in the current research and directions for future research .....	35
2.5 Summary .....	37
Chapter 3. Data and methodology .....	38
3.1 Introduction and context of data collection .....	38
3.2 Description of participants.....	39
3.3 Data used in this thesis .....	40
3.3.1 Oral interview data .....	40
3.3.2 Self-assessment data .....	44
3.3.3 Standardized tests.....	46
3.4 Analysis .....	49
3.5 Summary.....	51
Chapter 4. CAF analysis – results .....	52
4.1 Introduction .....	52
4.2 Complexity measures.....	52
4.2.1 Lexical variety.....	52
4.2.2 Lexical sophistication .....	57
4.2.3 Syntactic complexity .....	77
4.2.4 Syntactic variety.....	83
4.3 Accuracy measures .....	86
4.4 Fluency measures.....	86
4.4.1 Speech rate .....	88
4.4.2 Pauses .....	90
4.4.3 Self-repairs .....	93
4.4.4 Disfluency.....	95
4.5 Summary.....	97
Chapter 5. Discussion.....	98
5.1 Introduction .....	98

5.2	Comprehensive profiles: Change over time.....	98
5.3	Reflections on CAF analysis.....	106
5.3.1	Complexity .....	106
5.3.2	Accuracy .....	114
5.3.3	Fluency .....	116
5.4	Limitations.....	118
5.5	Implications and future work.....	121
5.5.1	Importance.....	121
5.5.2	Recommendations and caveats .....	124
5.5.3	Research needed.....	125
5.5.4	Extensions .....	127
5.6	Summary .....	128
Chapter 6.	Conclusion.....	129
References	.....	131
Appendix A – Excerpts from the CEFR for languages: Level descriptors	.....	141
Appendix B – Sample results of grammatical parsers applied to participant interview data	.....	144
B.1	ICTCLAS part-of-speech analysis .....	144
B.1.1	Input (arbitrary text from Elliott’s 2010 interview data) .....	144
B.1.2	ICTCLAS output.....	144
B.2	Stanford Parser grammatical analysis.....	145
B.2.1	Input (arbitrary text from Elliott’s 2010 interview data) .....	145
B.2.2	Stanford Parser output .....	145
B.2.2.3	Typed dependencies, collapsed .....	148
B.2.3	Input (same text from Elliott’s 2010 interview data, cleaned version) .....	149
B.2.4	Stanford Parser output .....	149

## List of Tables

Table 2-1: Comparison of assessment tools: CEFR, ACTFL, HSK, and TOCFL .....	9
Table 2-2: Complexity measures used in recent CAL studies .....	24
Table 2-3: Accuracy measures used in recent CAL studies .....	28
Table 2-4: Fluency measures used in recent CAL studies ( <i>continued on page 35</i> ) .....	32
Table 3-1: Participants' Chinese learning backgrounds .....	40
Table 3-2: Oral interview format in 2009 and 2010 .....	41
Table 3-3: Participants' results on HSK, TOP, and TOCFL standardized tests .....	47
Table 4-1: ICTCLAS-segmented 'outlier' tokens from interviews (not on the LCMC-5000 or HSK list) .....	65
Table 4-2: Participant use of the morpheme 到 <i>dào</i> .....	74
Table 4-3: Participant use of the morpheme 来 <i>lái</i> .....	74
Table 4-4: POS-tagging output from two grammatical parsers .....	79
Table 4-5: Participant use of the morpheme 过 <i>guò</i> .....	84
Table 4-6: Participant use of the morpheme 了 <i>le</i> .....	85
Table 4-7: Observed fluency measurements derived from interview speech samples .....	87
Table 4-8: Calculated fluency indices derived from observed fluency measurements .....	88
Table 5-1: Ella's CAF profile and changes over time .....	101
Table 5-2: Elliott's CAF profile and changes over time .....	102
Table 5-3: Patsy's CAF profile and changes over time .....	103
Table 5-4: Roma's CAF profile and changes over time .....	104
Table 5-5: Tim's CAF profile and changes over time .....	105
Table A-1: CEFR global level descriptions .....	141
Table A-2: CEFR spoken interaction level descriptions .....	142
Table A-3: CEFR spoken production level descriptions .....	143

## List of Figures

Figure 3-1: CEFR self-assessments for each participant in 2009 and 2010 .....	45
Figure 4-1: TTR vs. MATTR calculated for different window sizes of 20, 40, 60, 80, and 100 words. MATTR 100 is highlighted as being closest in discriminatory power to TTR. ....	53
Figure 4-2: MATTR with window size of 100 words calculated for different sample sizes .....	55
Figure 4-3: Percentage of words from interviews falling into different frequency ranges of Lancaster Corpus of Mandarin Chinese (LCMC) 5000-most-frequent token list (LCMC-5000) in 2009 and 2010 .....	59
Figure 4-4: Percentage of words from interviews falling into different HSK vocabulary levels in 2009 and 2010 .....	61
Figure 4-5: Mean word frequency of participant speech with reference to LCMC-5000.....	76
Figure 4-6: Speech rate fluency indices for each participant .....	89
Figure 4-7: Fluency indices: Pause counts .....	91
Figure 4-8: Fluency indices: Pause length and speech length .....	92
Figure 4-9: Fluency indices: Repair fluency.....	94
Figure 4-10: Fluency indices: Disfluency .....	96

## Acknowledgements

I would like to take this opportunity to thank Dr. Duanduan Li and Dr. Patricia Duff for their unwavering support, enthusiastic encouragement, and invaluable advice in pursuit of my studies and in life generally. The initial CRCCLE graduate student research team of Tim Anderson, Roma Illyckyj, Ella Van-Gaya (Lester), and Rachel Wang were also great friends and colleagues throughout our projects together; their intelligence, openness, and energy made this thesis possible. My thanks go especially to Rachel, who, on behalf of CRCLLE, worked tirelessly and creatively on the larger project, and on whose efforts this thesis rests in part.

Among the many inspirational researchers who I've had the honour of meeting or corresponding with, I'd like to single out Dr. Hongyin Tao (University of California, Los Angeles) for his keen teaching and encouragement, and Dr. Jun Da (Middle Tennessee State University) for his ideas, openness, and enthusiasm. Their help was instrumental in realizing this thesis. In addition, feedback from Dr. Edward McDonald from the University of New South Wales concerning some of the linguistic analysis was very helpful and much appreciated.

I also wish to recognize the graciousness, support, and good humour of the office staff of the Asian Studies Department at the University of British Columbia, and the department head, Dr. Ross King. I would not have found my way through my degree without their assistance and understanding, nor that of the librarians in the Asian Library or the logistical and academic advice provided by Dr. Josephine Chiu-Duke, Dr. Stefania Burk, Dr. Christina Laffin, and Dr. Sharalyn Orbaugh. Finally, Dr. Rebecca Chau and Dr. Sharalyn Orbaugh provided insightful comments on and suggestions for the thesis manuscript, and I'm indebted to them for their attention and guidance.

Last but most certainly not least, Si Nae Park's wonderful ideas, patient understanding, and steadfast support will always be remembered.



Finally, support for this research was obtained from a Social Sciences and Humanities Research Council grant to Drs. Duanduan Li and Patricia Duff, and from CRCLE.

## Chapter 1. Introduction

The field of Chinese applied linguistics, and particularly Chinese as a second or foreign language, is catching up rapidly to the advanced research status of languages such as English, French, and Spanish. China's cultural influence internationally has risen steadily alongside its economic power, and as it becomes a popular 'less commonly taught language' (and perhaps one day a 'commonly taught language') in North America and Europe studies involving learners of Chinese have played an increasing role in scholarship in applied linguistics. This thesis aims to raise awareness of an established subfield of applied linguistics, that of proficiency assessment through complexity, accuracy, and fluency (CAF) indices (see e.g. the special issue of *Applied Linguistics*: Hyland & Zuengler, 2009), in the context of Chinese learning, in hopes of promoting standardization in Chinese applied linguistics research and stimulating that research to make new observations about CAF studies extending beyond Chinese in the long term.

Chinese as an additional language (CAL) is the term used in this thesis to signify the study of Chinese as a non-native language, and encompasses the study of Chinese as a foreign language (CFL), Chinese as a second language (CSL), and Chinese as a heritage language (CHL). An increasing wave of CAL learners, both within China and abroad, has led to a boom in CAL pedagogy, research, and assessment (to validate candidates' proficiency for study or work). CAF studies, systematic investigation of which was popularized through task-based learning research (Skehan, 1996; Foster & Skehan, 1996), attempt to characterize discrete but interrelated elements of language production and performance such as lexical command, syntactic complexity, and speech rate, are important to all of these growing fields, and yet they have not been well studied in the CAL context to date, despite emphasis by researchers in the field of Second Language Acquisition (SLA) such as Tremblay (2011) that "documenting and controlling for L2

[second language] learners' proficiency in the target language should no longer be optional in experimental SLA research that seeks to explain the linguistic knowledge and behavior of L2 learners" (p. 364).

This study begins to address the gap in CAL research on CAF in learners' linguistic development by tentatively operationalizing some of the constructs proposed in the past, suggesting new ones where possible, and providing a concrete blueprint for researchers to make better use of computer technology in their investigations. Chapter 2 of this thesis reviews recent scholarship on CAF, with a focus on CAF in the CAL context, and provides definitions and an overview of CAF constructs which the rest of the thesis will build on. Chapter 3 describes the data collected for this thesis and the context it was collected in, then outlines the principles of the methodology employed to analyze it. Chapter 4 then details the different CAF constructs used to characterize the data, and finally Chapter 5 discusses in a broad way what can be learned from these CAF constructs, what to keep in mind when applying CAF methodology in the CAL context, and recommendations for future extensions to this work.

The following are the research questions which guided the work:

1. What techniques and tools are efficient and useful for characterizing linguistic patterns, sophistication, and variability in oral production data in Chinese as an Additional Language (CAL)?
2. Do these tools help to characterize meaningful differences in Chinese oral production and knowledge among learners:
  - representing a range of proficiency levels and previous formal language study?
  - in capturing differences over time, i.e. language development?
  - in comparing against benchmarks, or otherwise objectively determining proficiency?

3. What kinds of analysis can best be facilitated by computer tools such as those developed for the field of computational linguistics? What kinds of analysis might benefit from further developments in computer-assisted tools? What aspects (lexis/morphology, syntax, fluency, discourse) are not easily measured via computer-assisted tools, and could be investigated further in future research? What features need to be improved in the realm of freely available computational linguistic software for the purposes of describing oral language use and proficiency?
4. What are some of the pedagogical implications of the different tools and analyses available for the linguistic analysis of oral Chinese? In what ways can such analyses assist Chinese language instructors and learners, including those learning abroad in Chinese-speaking regions? What implications are there for textbook design and course delivery?
5. How are attempts to analyze and characterize the learning of oral Chinese different from discussions of oral proficiency development in other languages?

Cruickshank & Tsung (2010), in the conclusion to their recent edited volume on Chinese learning and teaching, provide a summary 'wish list' of research items for teaching and learning CAL; one of the wishes is for "applied linguistics research into second language development in Chinese," and another is "research into the development and validation of appropriate assessment instruments and tools" (p.223). As the authors note, these wishes may not be unrealistic given the recent surge in interest for CAL studies, and to that end it is my hope that this thesis represents one stepping stone on the path to better frameworks for CAL development and assessment research.

## **Chapter 2. Research in proficiency assessment through complexity, accuracy, and fluency measures**

### **2.1 Introduction**

This study focuses on measuring learners' oral Chinese proficiency, where:

- assessment is performed on a post-hoc basis (after data have been collected and without resort to clarifying by soliciting new data);
- language is transcribed or comes in written form; and,
- data are not generated via a discrete-point language test developed in advance (that is, there is no 'answer key' to check accuracy or correctness against, verify obvious grammatically correct answers, or determine if the use or avoidance of certain target language is considered successful).

Therefore, the primary areas of previous study that inform the analyses developed for this study are those concerning the measurement of complexity, accuracy, and fluency in language production; studies investigating the application of standardized or carefully constructed norm-referenced tests are less applicable in this exploratory context.

### **2.2 Measuring proficiency**

There are many possible ways to independently assess proficiency in language: for example, specific criterion-referenced and achievement tests built to assess knowledge of certain grammatical structures or vocabulary, often based on previous material established by a particular instructor and presented to his or her pupil(s); standardized tests, developed by boards of language professionals and

psychometricians; and qualitative assessment, sometimes based on a rubric, of less-tester-controlled learner output, such as essays or dialogue (for an overview of recent use of assessment methodologies, see Bachman, 1991; Bachman, 2000; Tremblay, 2011). In applied linguistics research, in lieu of independent tests, sometimes stand-ins are used to attempt to indirectly imply current proficiency, such as years of second language (L2) study, previously awarded proficiency scores, or self-ratings (Tremblay, 2011).

What does it mean to be proficient? Obviously, grammatical accuracy is not the only aspect of communication that interlocutors notice. Although a full discussion of the various frameworks proposed in the field of linguistics to define proficiency is outside the scope of this thesis, the concept of ‘communicative competence’ is helpful in determining what proficiencies are brought to bear by a language user. Canale & Swain’s (1980) description of communicative competency can be summarized as a combination of grammatical competence (i.e. knowledge of and accuracy in using lexicon, syntax, morphology, and phonology), sociolinguistic competence (i.e. awareness of discourse ‘rules’ and situational and linguistic appropriateness), and strategic competence (i.e. strategies employed to accommodate weakness in other competencies or ambiguity in discourse). Bachman & Palmer (1982) proposed different sub-components of communicative competency (grammatical, pragmatic, and sociolinguistic), but the idea remains that proficiency is not a unitary construct. Though perhaps it is not as simplistic a case as saying each factor is exactly equivalent in importance in every situation, proficiency, and full assessment thereof, should take into account not just the accuracy of the communication but also its appropriateness and effects in a particular context, and the context of production and data collection may determine what mixture of proficiencies are actually being tested at any given time.

What should a good assessment be sure of? Malone (2011) presents key concepts in assessment: validity (fidelity in measuring target aspects of proficiency and the appropriateness of the measure for a particular test audience or purpose); reliability (consistency of test results across different times, ver-

sions, and assessors); practicality (feasibility of implementing a test in terms of resources); and impact (how a test affects educators, test-takers, and the community). When preparing assessments or choosing among those available, these will be key decision items in the evaluation process; each assessment technique presents a different mixture of strengths across these four factors.

For instance, a particularly salient point for the purposes of the current thesis is validity and reliability of oral proficiency testing: Can judgments of oral proficiency consistently and accurately predict overall proficiency (i.e., across multiple skill areas: reading, writing, listening, interaction, as well as speaking)? As pointed out by many researchers, different language learners exhibit different mixtures of confidence and competence and different strengths in particular areas of proficiency. For instance, Li & Duff (2008) note that heritage learners initially often show high degrees of oral fluency, but limited vocabulary, only modest awareness or command of different registers, and relatively underdeveloped literacy skills. Even amongst non-heritage learners there are differences between sojourners, who may learn informally in a second language environment, and students in academia, who may learn in a structured, foreign language classroom environment. This question is again much larger than the scope of this thesis, and so the focus of this study will be on how to operationalize various techniques of characterizing proficiency (Chapters 3 and 4), followed by a discussion of their possible strengths and weaknesses with respect to the types of proficiency they measure and how well they succeed.

### **2.2.1 Standardized assessments in the CAL context**

In the field of CAL studies, the primary internationally recognized tests in recent times are the Hanyu Shuiping Kaoshi (HSK), administered by Hanban (Beijing) internationally, and the Test of Chinese as a Foreign Language (TOCFL), developed by the Steering Committee of the Test of Chinese Proficiency-Huayu (SC-TOP) and used as the HSK equivalent in Taiwan. There are also more specific tests for different purposes (e.g. Hanban's Business Chinese Test, or BCT). A great deal of research has aimed at devel-

oping these tests into robust and useful gauges of non-native Chinese language capability which can serve as standards both domestically and internationally (for example, see Luo et al., 2011; Wang, 2011).<sup>1</sup>

The pre-2008 HSK categorized learners into broad ranges, each with a subset of numbered micro-levels. The current HSK indexes learner proficiency using a numbered scale from 1 to 6, with 6 being the most advanced. Prior to 2010, TOCFL was known as the Test of Proficiency in Hanyu (TOP-Huayu, or TOP). The TOP-Huayu levels were similar in format to the old HSK levels, and the newer TOCFL proficiency levels map directly onto the old TOP-Huayu levels with new names.

In North America, a test called the Oral Proficiency Interview (OPI), part of the American Council on the Teaching of Foreign Languages (ACTFL) framework for assessment, is a common way to assess oral proficiency in Chinese, and learner proficiency is characterized using the ACTFL scale.<sup>2</sup> It is not used in this thesis other than as a rough reference point for comparison (see Table 2-1, below).

In Europe, and now some other regions as well, yet another standardized assessment, or rather assessment framework, is the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001).<sup>3</sup> It provides robust and concrete guidelines for developing assessment mechanisms and characterizing learner proficiency which have been used widely in European language teaching for the past decade. Many European countries' national standardized language tests have been aligned with

---

<sup>1</sup> For more information on research into development and validation of these standardized tests, see [http://www.hsk.org.cn/hskstuff/all\\_paper.aspx](http://www.hsk.org.cn/hskstuff/all_paper.aspx) (for HSK) and <http://www.sc-top.org.tw/chinese/publication.php> (for TOCFL).

<sup>2</sup> For more information on the ACTFL OPI, see <http://actfl.org>.

<sup>3</sup> For more information on the CEFR and how it can be used to calibrate and guide language learning, see the language policy website of the Council of Europe, <http://www.coe.int/lang>.



the CEFR scale, and proficiency guidelines for many world languages have been adopted and refined.<sup>4</sup>

One benefit of the CEFR is a comprehensive set of resources available in many languages for learner self-assessments, based on ‘can-do’ statements that allow learners to reflect on their own capabilities in real language situations and use those observations to find a CEFR rating commensurate with their abilities; it also gives learners guidance on advancement by outlining in detail the specific tasks a learner at a higher level is expected to perform.

The CEFR defines six general levels, also known as the global scale, from A1 (‘Breakthrough,’ the lowest level of proficiency) to C2 (‘Mastery,’ the highest level of proficiency in the framework). Global levels are therefore quite broad; for example, the spoken interaction sub-skill description for level A2 (Council of Europe, 2001) includes the ‘can-do’ statements “I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself.” In contrast, the spoken interaction sub-skill description for the next-higher level, B1, is as follows: “I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).” For detailed descriptions of all sub-skills at each global level, along with the overarching global level descriptions, see Shneider & Lenz (2003), or a reproduced summary of the spoken interaction and spoken production sub-skills in Appendix A of this thesis.

Table 2-1 shows how the CEFR, ACTFL, HSK, and TOCFL scales are related.

---

<sup>4</sup> For an example of a project in the Canadian context to apply the CEFR regionally to specific languages, see BC Ministry of Education (2010).

**Table 2-1: Comparison of assessment tools: CEFR, ACTFL, HSK, and TOCFL<sup>5</sup>**

CEFR	ACTFL	HSK (old)	HSK (current)	TOP-Huayu (old)	TOCFL (current)
A1	NL - NH	<i>(data unavailable)</i>	1	–	<i>(data unavailable)</i>
A2	NM – IH	<i>(data unavailable)</i>	2	Beginner	Beginner (2)
B1	IL – AL	<i>(data unavailable)</i>	3	Basic (1 – 2)	Learner (3)
B2	IH – AH	Elementary (3 – 5)	4	Intermediate (3 – 4)	Superior (4)
C1	AM – S	Intermediate (6 – 8)	5	Advanced (5 – 7)	Master (5)
C2	AH – D	Advanced (9 – 11)	6	–	–

*For ACTFL, Novice Low, Mid, High = NL, NM, NH; Intermediate Low, Mid, High = IL, IM, IH, Advanced Low, Mid, High = AL, AM, AH; Superior = S; Distinguished = D. TOCFL (current) levels show the pre-2012 names, with the 2012 numbered levels in parentheses.*

## 2.2.2 Specialized assessment methods

Other research has been conducted to look at ways of assessing learner proficiency with respect to specific types of linguistic knowledge. Tremblay (2011) briefly reviews many competing (or coexisting, and sometimes blithely incomparable) methods of proficiency assessment used in prominent SLA research between the years 2000-2008, including interactive or administered methods such as customized or partial standardized tests, cloze tests, and oral interviews, and other descriptive measures or information, such as length of study, length of exposure, classroom level, pre-existing proficiency test scores, and self-ratings. Tremblay aptly voices the commonly-heard caution: “...comparisons between studies—at least those that investigate the same target language—would be more amenable if there were at

---

<sup>5</sup> HSK (old) levels and names were derived from UBC (2006).

HSK (current) – CEFR alignment is from Hanban (n.d.). HSK (old) – HSK (current) alignment is from Hanban (2011). Note that HSK (old) Basic levels 1-3 are not shown as they were not aligned with the CEFR according to Hanban (2011).

TOCFL – CEFR alignment is from Chang (2011). TOP – TOCFL alignment is from SC-TOP (2010); TOCFL level numbers are from a news posting on <http://www.sc-top.org.tw/> dated January 13, 2012 and retrieved on May 4, 2012.

CEFR – ACTFL alignments represent a wide range of opinions from various scholars; these ranges are comprised of proposed equivalencies derived from or quoted in Chang (2011), Martínez Baztán (2008), and Vandergrift (2006).

least some consensus on acceptable procedures that researchers could use to estimate proficiency and more consistency in the characterization of these procedures” (p. 341).

### **2.2.3 Issues with standardized and specialized assessment methods**

The standardized tests described above are constrained by certain limitations. The HSK and TOP can only be taken at certain times and in certain places, and the assessments are also done by their respective governing bodies according to answer keys that are not made public. The OPI must be administered by a trained assessor. A tool developed at the Center for Applied Linguistics called the SOPI (‘simulated OPI’) is a more streamlined and flexible approach; it uses a question booklet and requires the learner to record oral responses to prompts in the booklet. A computerized OPI (COPI) for Mandarin Chinese is also currently being developed by the Center for Applied Linguistics.<sup>6</sup> The SOPI and COPI still require an authorized assessor to follow a set of guidelines in gauging the learner’s oral proficiency, which is scored on the ACTFL scale.

Specialized tests, very useful for carefully observing linguistic phenomena in SLA studies or for testing discrete types of linguistic knowledge in language courses where the teacher doesn’t have recourse to a good standardized solution, are difficult to compare with each other. Specialized tests may only be able to categorize learners based on specific scales, which are hard to align and equate with more general encapsulations of learner proficiency. Typically, specialized tests are created and administered for one specific group of learners, requiring assessor time (to develop and score the assessment)

---

<sup>6</sup> For more information on research and developments in Chinese testing under the SOPI and COPI frameworks, please see:

SOPI: <http://www.cal.org/topics/ta/sopi.html>

SOPI: <http://www.cal.org/resources/digest/0014simulated.html>

COPI: <http://www.cal.org/about/calnews/archive/111011.html>

and expertise (to design an assessment that does indeed measure the target skill in a fair and consistent way).

#### 2.2.4 Characterizing proficiency: Complexity, accuracy, and fluency (CAF) studies

One form of proficiency assessment that has been used for many years and lends itself well to automated or semi-automated, repeatable, and quantifiable analysis is the set of constructs known as **complexity, accuracy, and fluency**, often abbreviated **CAF** in SLA literature (Housen & Kuiken, 2009). These three constructs have been studied in recent years in order to try to characterize language use in various stages and situations, with a view to linking the various operationalized indices that correspond with the constructs to models of the underlying psychological systems brought to bear in creating language, and of course to linking them with other proficiency assessments in order to verify the validity of those assessments and provide a robust framework for describing e.g. developmental phases or changes in speech (for example, see Iwashita et al., 2008; Kuiken et al., 2010).

Ellis (2003) described these three dimensions of proficiency as follows:

- **Complexity**, “The extent to which the language produced in performing a task is elaborate and varied” (p. 340);
- **Accuracy**: “The extent to which the language produced in performing a task conforms to native speaker norms” (p. 339); and,
- **Fluency** : “The extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation” (p. 342).

Complexity and accuracy are increasingly believed to be indicative of underlying interlanguage knowledge, and fluency to be related to control of L2 knowledge (Housen & Kuiken, 2009). Researchers believe them to be interrelated, and it is posited that they interact in complex ways such that learner

cognitive resources focused on one of the three may mean a drop in one or more of the others; indeed, much of the work examining the validity of the constructs and hunting for clues to our underlying psycholinguistic processes use task-based language learning research, some of which has investigated the ‘trade-off’ when a task requires extra attention in any given CAF dimension. As an example, Yuan (2009) points out that errors may be more common for learners attempting to use complicated language; accuracy is in this case inversely proportional to complexity.

Each of these CAF constructs will be briefly described as follows, followed by their application to date in the field of CAL studies, in section 2.3.

#### **2.2.4.1 Complexity**

Useful measures of complexity are still under active investigation, as it is not necessarily a single factor connecting directly to language proficiency; indeed it is currently understood as deriving from two main features, **cognitive complexity**, “defined from the perspective of the L2 learner-user” (i.e. complexity in the processes of acquisition and performance, also encompassing learner variables such as memory, motivation, and language background), and **linguistic complexity**, “defined from the perspective of the L2 system or the L2 features” (i.e. structural or functional complexity of elements of the target language system, or depth and richness of a learner’s developed interlanguage system) (Housen & Kuiken, 2009, p.463). Recent studies seem to bear out the logical idea that order of acquisition tendencies for different first language (L1) – second language (L2) pairs is indicative of cognitive complexity, such that for L2 speakers, the more cognitively complex an (attempted) utterance, the more fluency and accuracy may be negatively impacted in the produced language (see for example Kuiken & Vedder, 2008; Salamoura & Saville, 2010). For example, given L1 English and L2 Chinese typological differences, acquisition of word order in *ba*-construction sentences (which typically highlight an object and the result of an action upon that object) could be considered more cognitively complex than acquisition of the de-

clarative sentence word order (subject-verb-object) for English learners of Chinese, as the latter trait is shared by both languages and therefore posited to be less difficult to learn and produce. As for linguistic complexity, “when considered at the level of the learner’s interlanguage system, [it] has been commonly interpreted as the size, elaborateness, richness, and diversity of the learner’s linguistic L2 system” (p.464). Linguistic complexity is believed to give an indication of the state of a learner’s developing knowledge and command over the target language. Finally, linguistic features themselves are said to possess **structural complexity**, i.e. inherent formal or functional complexities; an English example might be that the {noun phrase – verb} construction “he went” is of lower structural complexity than the {noun phrase – verb – noun phrase – adjectival phrase} construction “he painted the car red” (Salamoura & Saville, 2010, p. 116).

Complexity can be broadly categorized as follows:

**Lexical variety or lexical diversity:** The classic measure developed to indicate lexical variety in language is *type-token ratio (TTR)*, which looks at the ratio of unique words (‘types’) to total words (‘tokens’). Operationalization seems straightforward but does require some decisiveness; for instance, in some languages, like English, the *type* can indicate a single (e.g. root) form of a set of words different in form but deemed equivalent in meaning (e.g. *do, does*), and in some cases two tokens with the same form may be deemed to be separate types as well (e.g. polysemous words such as *to try: try it out* vs. *try one’s patience*). Iwashita (2010), however, points out that some studies have shown *TTR* to be an imperfect demonstration of lexical variety, and thus some techniques have been developed to improve its applicability across a range of proficiency levels. To that end, *TTR, mean segmental type token ratio (MSTTR)*, the average of *TTR* calculations for small sequential lengths or blocks of texts (called ‘windows’ due to the concept of a ‘frame’ outlining a block of text to be analyzed using an algorithm), and *D value* (or simply, *D*), obtained by curve-fitting mathematical operations, are both well described by Malvern & Richards (2002), with the latter showing the greatest general applicability and validity across multiple

sample types. Covington & McFall (2010) discuss other possible alternatives to *TTR*, culminating in the suggested use of *moving-average type-token ratio (MATTR)*, an improvement on *MSTTR* and on *D*.

*MATTR* is introduced as a way to examine the change in style of a text over its length, and uses the concept of a moving, overlapping window; although different values result from different window sizes used in the calculation, thanks to this moving window, it can account for all morphemes in a text of any length (the *MSTTR* algorithm requires that some morphemes be jettisoned if the text length is not a multiple of the block size used).<sup>7</sup>

Despite the concern over current methods of quantitatively assessing lexical variety, Skehan (2009) argues that measures of lexical performance are vital additions to CAF measurements, noting its importance to both pre-performance analysis (how complex is language input?) and performance analysis (how complex is learner output?).

**Lexical sophistication:** Attempts to measure lexical sophistication include *TTRs* in which the type parameter is refined to some interesting subset of language, for instance words of a certain difficulty (conforming to a word list), parts of speech, and so forth. Another way to examine sophisticated lexical use is to look at the produced words' frequencies in established reference corpora, equating lower-frequency words with a higher level of lexical sophistication.

**Syntactic complexity or syntactic maturity:** Researchers have used different speech units to assess syntax complexity, but generally in English they have looked at the amount of subordination and

---

<sup>7</sup> The following is a simple illustration of the *MATTR* algorithm's moving window concept. *TTR* calculations are made for each block of text of length equal to the window size; the first window starts with the first word, and the window frame starts (and ends) one word further into the text on each iteration of the algorithm, until the final word of the text is included in a window frame. For example, the text "I love the Chinese language" is comprised of five words; when analyzed using *MATTR* with a window size of three words, the following three analytical samples are produced as the window 'moves' through the text: "I love the, love the Chinese, the Chinese language." A *TTR* would be calculated for each measurement (for a total of three measurements), and the average is the *MATTR*.

coordination present in a single sentence or cohesive speech unit. In written works, text is generally analyzed by dividing it into sentences, clauses, or *T-units* ('terminal units'). T-units are defined in Yuan (2009, p. 125) as "'the shortest units into which a piece of discourse can be cut without leaving any sentence fragments as residue' (Hunt, 1970, p. 188)" or "'one main clause with all subordinate clauses attached to it' (Hunt, 1965, p. 20)." Complexity is often an investigation of T-unit length, clauses per T-unit, and in second language studies (usually of writing) commonly includes percentage error-free T-units as well. Speech presents further difficulty in terms of clear units of division, and researchers have proposed other measures of cohesive speech unit, such as the utterance, c-unit ('communication unit'), and AS-unit ('Analysis of Speech unit') (Crookes, 1990; Norris & Ortega, 2009).

In the field of CAL acquisition, alternatives to traditional complexity measures have been proposed considering the typological properties that differentiate Chinese from the bulk of the (predominantly Western) languages that have informed SLA CAF studies, which will be covered in further detail in 2.3.1.

**Syntactic variety:** This element of complexity is used to indicate the variety and sophistication of grammatical structures, including the range of structures produced and variation in or depth of use of specific structures. In languages with inflectional and derivational morphology, such as English and Spanish, the variety of morphological changes exhibited has been considered an indication of syntactic variety, but for languages like Chinese that do not exhibit verb inflection, other measures have been proposed, such as assessing the variety of pedagogically salient sentence patterns used. For example, in French a researcher might look at morphological variety in verb use as an indication of control over tense, modality, and voice; in Chinese one might instead look at the variety of use of *ba*-constructions, *bei*-constructions (explicit passive marking), resultative verb constructions, and so forth.



#### 2.2.4.2 Accuracy

Researchers typically define accuracy indices as measures of how closely language adheres to a baseline of expectations or rules, typically prescriptive, native speaker (NS) norms. Indices can be used to measure **global accuracy**, i.e. “identifying any and all types of errors” (Iwashita et al., 2008, p.35) e.g. over a length of language output, or **local accuracy**, using specific error types as an index. Global measures can be difficult to code consistently, whereas local measures are only effective as an approximation to global measures, but are often quicker and easier to identify, and are thus frequently used in studying specific aspects or linguistic features in SLA studies.<sup>8</sup> However, as Yuan (2009, p.118) cautions, “local measures can only be used when the data are collected from form-focused tasks such as error correction or sentence translation, or from the data with a meaningful quantity of the linguistic form in question.”

Accuracy can be broadly classified into one of the following types:

**Error frequency:** Usually expressed as a percentage (e.g. sentences without errors as a proportion of total sentences, or error-free clauses or T-units as a proportion of total clauses or T-units) or an **error density** index (i.e. errors per standard length of speech/text), error frequency is the most common way that researchers have analyzed learner accuracy. As Yuan (2009) points out, such error density measures can give more precise indications of high error frequency as error-free percentage measures often do not distinguish sentences with multiple errors.

**Self-repairs:** Repair measures take into account **false starts** and **repetitions** (Housen & Kuiken, 2009); false starts are often called **self-repairs** (as opposed to corrections from interlocutors). Although

---

<sup>8</sup> For further discussion of theoretical and operationalization issues with accuracy indices, including the definitions of an ‘error’ and what might constitute target-like usage, see Housen and Kuiken (2009, p.463).

commonly associated with disfluency, in that self-repair behaviour can slow delivery of a message, a high incidence of self-repair is believed to indicate a greater awareness of form and clarity, and thus self-repair frequency is considered an indirect way of measuring accuracy.

**Error type** and **error gravity**: One issue with indices of global error frequency and self-repair frequency is that error variety and gravity is not taken into account. Some errors have a more serious impact on meaning or sociocultural intent than others, and in certain circumstances an otherwise proficient speaker might exhibit a high frequency of a particular type of error (perhaps one associated with sophisticated language use) which might skew the perception of global error frequency measures for that speaker, despite target-like use of many other linguistic features. As Yuan (2009) points out, however, “up till now few studies have employed measures of error gravity and error type possibly due to the inherent complexity in setting up specific indexes since too many parameters have to be taken into consideration” (p.119).

#### **2.2.4.3 Fluency**

Fluency indices have primarily been studied to “determine which quantifiable linguistic phenomena contribute to perceptions of fluency in L2 speech” (Housen & Kuiken, 2009, p. 463), and can be subdivided into dimensions of **speed fluency**, **breakdown fluency**, and **repair fluency**.

**Speech fluency**: This rather broad-sounding term is a catchall for speech production and speech rate measures. **Speech rate** is commonly calculated in syllables per second or minute, and typically these calculations take into account disfluencies (see breakdown fluency, below). An example of this is the **Menhert production rate**, or **quality production rate**, which excludes repair and repetition (see repair fluency, below) when calculating rate of speech (Yuan, 2009).

**Breakdown fluency**: Pauses are integral to both non-native speaker (NNS) and NS discourse, but breakdown measures try to capture **disfluent** or non-target-like pauses. These disfluent pauses, also

known as **false pauses**, are posited to be disruptions due to linguistic proficiency difficulties rather than rhetorical effect or the need for further thought. Obviously, it can be difficult to parse out which pauses in discourse are disfluent and which are not; researchers have proposed various ways to operationalize this parameter, including factors of length (often they decide on a threshold length above which a pause is considered disfluent), position (within or between clauses, for instance), and classification (whether a pause is **filled**, i.e. whether sounds or discourse features like ‘*um*’ ‘*ah*’ ‘*like*’ etc. are used to pause between fluent segments, or **silent**).

**Repair fluency:** This type of measure characterizes the proportion of speech time or speech produced that involves self-repair behaviour (see 2.2.4.2 for a description of self-repairs). Typically repair fluency is calculated as a density, e.g. repairs per x words, or as a percentage of total speech, e.g. repair syllables / total syllables. Self-repairs and repetitions are often removed from consideration when performing analysis of syntactic and lexical complexity/variety, and are sometimes called **pruned** syllables or pruned speech in this case.

## 2.3 CAF in the CAL context

Most SLA research on CAF published in Chinese (mainly that produced in China) focuses on English as a second or foreign language, rather than CAL, likely due to the presence of EFL instruction as a mandatory part of China’s curriculum, and the breadth and history of EFL/ESL studies to compare results with. There are many Chinese-language studies of CAL learners that employ error analysis techniques (e.g. Han, 2003; Li et al., 2007); these could be useful for developing automated or semi-automated ways of determining accuracy in CAL production, but are not the focus of this thesis and will not be dealt with further except as ideas for improving the set of constructs used for assessing CAL proficiency (see Chapter 5).

Few studies have looked at CAF in the context of CAL. Yuan (2009) explains that CAL studies in SLA have tended to look at accuracy of use of certain grammatical structures (local accuracy measures), e.g. target-like use of *ba*, *le*, *ma*, *guo*, *zhe*, and word order, and not global accuracy measures or the other dimensions of complexity and fluency. In Yuan's words, "no attempts have been found to integrate [CAF indices in CAL studies] together within the framework established in general L2, and develop similar measures that can be used to code, analyze, and evaluate L2 Chinese learners' performance" (p. 110). Indeed, it is important to note that typological differences make the application of non-Chinese-target-language CAF research problematic in the CAL context: word boundary decisions, word order differences, and topic-prominence (versus subject-prominence) all affect the way in which CAF measures are operationalized. These issues will be dealt with as the various indices are introduced or discussed in this thesis.

Some recent examples of CAL studies using CAF methodology are briefly described here, and the conclusions drawn from them on the use of specific indices relevant to CAL studies will be discussed in 2.3.1.

Shen (2005) examined the relationship between lexical sophistication (word frequencies), learner vocabulary knowledge (what characters were 'known' by a reader), and text complexity ('sub-units' per sentence, sentence length, and so forth; see below for brief critique), on the one hand, and CAL reading comprehension, on the other. The data under examination were CFL textbook reading passages. Interestingly, in the Chinese context, Shen proposed the use of word frequency measures as an indirect index of the difficulty of grammatical structures, as "there is no extant inventory that provides the ranking scale for the difficulty level of [Chinese] grammar structures" (p.7); an example of this logic was that the *ba*-construction was posited to be 'easier' than the *bei*-construction due to their relative frequencies as given by a reference frequency dictionary. (Shen cautioned that further studies were necessary to determine if this was indeed a reliable approximation.) Shen's 'sub-unit' measures, based on orthographic boundaries and grammatical categories and described in some detail, were nevertheless not

clearly predicated on a standardized system of classification or detailed to a degree that would make them reproducible by other researchers.

Jin (2007) looked at the appropriateness of various indices to measure complexity in CAL, comparing language produced by L1-English-speaking learners of Chinese at various levels (delineated by length of study) to that produced by native speakers of Chinese. Jin argues the case for recasting syntactic complexity as **syntactic maturity**, since the concept of mature (NS-like) language varies for different language types; Jin found that the ‘complexity’ of Chinese NS speech was actually lower by some measures than NNS speech, due to characteristics of Chinese that Jin’s English-background test subjects did not exhibit in full, such as unmarked passives, omission of determiners and repetitive nouns, and reduced conjunction use indicative of topic-comment constructions. The primary conclusions of the study were that a balanced assessment of syntactic maturity in Chinese speech should take into account use of *zero elements* in multiple positions, *covert conjunctions*, *topic-comment* constructions, and *topic chaining*. The general findings, that NS speech can be ‘less complex’ than NNS speech by some common measures, e.g. T-unit length, were confirmed by Yuan (2009), discussed shortly.

Guo (2007) investigated the fluency construct in the context of CAL by analysing recordings of 30 individuals’ answers on the oral portion of the Advanced-level HSK (the ‘old HSK,’ see 2.2.1), manually transcribing the recordings, calculating 11 indices, and checking their validity. Guo’s calculations were described in sentences, but not elucidated with equations or consistent terminology; for instance, Guo did not specify the providence of all parameters used in the investigation, including 0.3s as a non-target-like pause, and conflated T-units with ‘sentences’, a difficult concept to define in speech texts. The study calculated a ‘fluency index’ for each of the 30 samples and for the purposes of horizontal comparison. The author assessed the set of measurements relating to time (speech rate, pauses) to be most indicative of fluency, but found little correlation with HSK oral test results, which the study attributed to fluen-

cy being only a component of oral proficiency. The author further judged the relatively subjective measures of ‘accuracy’ and ‘effectiveness’ used in the study to be unhelpful as indexes of fluency.

Another example of a study with nebulous methodology is Sun (2008), wherein 48 CSL students were split into three different groups to see the effect of task preparation on CAF measures (i.e., oral expression performance). The participants were pre-tested using the HSK and an oral test to ensure they were all roughly the same proficiency, but no details were given on the HSK scores or the makeup of the oral test. Although Sun described using familiar CAF measures such as number of repetitions, number of self-repairs, rate of speech, number of sentences with errors, sentence complexity, sentence type, and lexical diversity, these measures were not standardized,<sup>9</sup> and used an unexplained rating scale.

Yuan (2009) took stock of some of the major findings from CAF studies in CAL acquisition research in the preceding decade, employing a set of these indices in an exploratory investigation of the differences that CAF measurements show between NS speech and CAL interlanguage. Yuan used a short excerpt of an intermediate CAL learner’s in-class speech as a test bed to illustrate how various CAF indices would apply to Chinese speech; the speech sample was transcribed, translated into English, then presented to a Chinese native speaker to in turn orally produce the same ideas in Chinese; Yuan finally transcribed the NS speech, using the transcript of the NNS speech and the transcript of the native speaker’s version as two comparable samples. While this method introduces many issues (very little data analyzed and presented to the reader, multiple forms of oral-to-written conversion without systematic explanation of transcription or elicitation procedures, and unqualified consultation of “third party

---

<sup>9</sup> One example of questionable parameter choice in Sun (2008) was setting 150-160 syllables per minute as the ‘normal speech’ standard, then docking marks from participants in an unspecified “concrete” way for slower speech (p. 90). This normal speech standard was apparently based on expectations of NS fluency, but the factors that informed choice of this range were not described, thus issues such as the contexts in which NS speech could be expected to meet those expectations and whether or not the NNS samples used for the study conformed to the same contexts were not addressed.

opinion” (p. 113), to name a few), the aim was to illustrate, non-conclusively, how CAF measures might be derived from data, concretizing the operationalization of some indices discussed in the literature review. Moreover, Yuan’s excellent summary of recent studies of CAF measures in the CAL context was clear, concise, and thorough, and is thus relied on to a great extent in this thesis. Yuan proposed further investigation of some promising measures, and cautioned that some obvious or common measures, such as T-unit length and ratio, verb form ratios, and inflection-based accuracy indices, were not conclusively helpful in characterizing Chinese proficiency.

Yuan (2010) examined Chinese L2 narrative writing produced by 42 advanced-low university students, split into three equal groups, to see how variations in task (e.g. providing guidance or imposing structural guidelines) affected linguistic performance. Yuan cautioned that measurement was based on constructs originally developed for analysis of English and other European languages, used “since there were no ready indices for Chinese when the study was conducted” (p. 85), but believed this was a limitation of the study, and subsequently invested efforts in developing measuring indices specifically for L2 Chinese (see Yuan, 2009, and discussed in section 2.3).<sup>10</sup> The study found that when guidelines were imposed on the types of grammatical structures learners were expected to use, the learners almost universally scored worse on CAF measures than the group of learners given no guidelines, with the exception of lexical sophistication. For instance, words produced per minute, which was equated with fluency in this study, dropped in the focus-on-form group. Yuan thus believed that “if [the learners] choose to or are forced to attend to one aspect, for example, accuracy, their performance may suffer in other areas, such as fluency and/or complexity” (p.84), i.e. the so-called ‘trade-off effect’ (see section 2.2.4).

---

<sup>10</sup> It must be noted that the discussion presented in Yuan (2009) and Yuan (2010) did not fully explain why measurements used in those studies might be more or less appropriate in a Chinese-language context.

Yuan (2009) discusses other examples of CAF analysis found in CAL literature, and as Yuan's measurements in that study incorporated assessment of those examples, they will not be explored in detail here. An exhaustive list of the various indices that have been used to measure CAF and their relative strengths and weaknesses is outside the scope of this thesis. For the purposes of CAF methodology applicability to CAL studies, however, the primary measurements that have been used to investigate L1 Chinese and L2 Chinese in existing research will be briefly discussed.

Finally, some graduate theses have been written examining CAL or L1 Chinese acquisition employing CAF methodologies (see for example Liu, 2007). For the purposes of the present study, only research summarized in journal articles was considered, as it has been peer reviewed and is widely accessible.

### **2.3.1 CAF measures in the CAL context: Complexity**

Table 2-2 lists the main complexity measures suggested or investigated in the field of CAL studies, along with variations on the operationalization parameters or interpretation of those measures.



**Table 2-2: Complexity measures used in recent CAL studies**

Index / construct / measurement	Variation	Notes on calculation	Recent studies			
			Yuan 2009	Yuan 2010	Jin 2007	Shen 2005
Lexical variety or lexical diversity	<i>type/token ratio (word variety)</i>	(number of different words) / (total words)		✓		
Lexical sophistication	<i>sophisticated word type/token</i>	e.g. (HSK words beyond B level) / (total words)	✓	✓		
Lexical sophistication	<i>mean word frequency</i>	mean of word frequency of entire text				✓
Lexical sophistication	<i>Word frequency range</i>	lowest and highest word frequency				✓
Syntactic complexity ('maturity')	<i>Clause density</i>	(number of clauses) / (T-unit)	✓		✓	✓
Syntactic complexity ('maturity')	<i>MLTC</i>	mean length of terminal topic-comment unit (TTCU)			✓	
Syntactic complexity ('maturity')	<i>C/TTCU</i>	(number of clauses) / (TTCU)			✓	

Some important typological issues that influence the operationalization of CAF measurements of complexity are as follows:

- **Word boundary:** As Yuan (2009) points out, “boundaries of Chinese words are not always clear” (p.113). Because Chinese text is not written with spaces between characters (morphemes), word segmentation is a non-trivial process, and hence tools to do such work make decisions based on standardized grammars and grammar trees and statistical models deriving from reference corpora. The ‘words’ such segmented output consist of are perhaps better thought of as ‘tokens,’ or discrete bundles of morphemes representing a single idea, which if divided would no longer represent that idea; this can make it difficult to compare such tokens with ‘words’ as defined in dictionaries and pedagogy.
- **Morphology:** Chinese verbs do not conjugate or change form the way verbs do in many other languages; the relationships of the verb to tense, aspect, actor, and so forth are communicated via additional particles or words used to provide context to the verb. This means that analysis of syntactic variety, for one, based on languages of Indo-European roots often are not easily transferable to the Chinese context.
- **Topic-prominence:** The phenomena of topic-prominence and topic-chaining are integral to Chinese discourse, and mean that traditional studies of syntactic complexity through subordination, coordination, length of utterance or T-unit, etc. may not be appropriate to determine whether or not Chinese speech is ‘native-like.’

Details of the methods of measuring complexity depicted in Table 2-2 are discussed in greater detail below.

**Lexical variety:** Yuan (2010) calculated lexical variety using a basic TTR measure (number of different words divided by total words). Yuan (2009) cited Huang and Qian's (2003) study as performing a similar lexical variety calculation, and also using a slight variation they called "lexical density", which was basically a TTR wherein the 'type' parameter excluded prepositions. Neither Yuan's 2010 study nor Huang and Qian's 2003 study found any significant differences for TTR across task condition groups or after one month's study time, respectively.

Yuan (2009) noted that for the small NS-NNS comparison experiment in that paper, the native speaker text showed a much higher lexical variety as measured by TTR. However, Yuan echoed the concerns summarized in section 2.2.4.1 that this type of measure may not be valid with longer texts, and that MSTTR might be a better way of investigating lexical variety.

**Lexical sophistication:** In CAL studies lexical sophistication is typically indexed with a sophisticated TTR measure, made easier by the ready availability of HSK vocabulary lists for different proficiency levels. These lists are generated by the HSK authority but their exact relationship to the HSK exam is not well described by Hanban.<sup>11</sup>

Yuan (2009) described a few previous studies that had used such a lexical sophistication measure (Huang and Qian, 2003; Shen, 2005; Yu, 2002), and one of those studies (Huang and Qian, 2003) did find a slight increase in lexical sophistication after one month's study. Yuan's sample NS-NNS comparison distinguished HSK levels 1, 2+, and 'special' vocabulary (not well explained in the text), and commented that the native speaker exhibited far more sophisticated lexical use. Yuan (2010) used an HSK A+

---

<sup>11</sup> Prior to 2009, when the HSK was revised (see 2.2.1), there were four levels of vocabulary, A to D, with A being the most basic level; as the level increased, the list for that level included a larger amount of vocabulary than the previous level, and of a generally higher sophistication. After the 2009 HSK revision, the new vocabulary list was split into six levels, with level 1 being the most basic.

TTR (words outside of the HSK A-level word list divided by total words), and found that the ‘focus on form’ experimental group did use a greater number of HSK A+ words.

**Syntactic complexity:** According to Jin (2007), there has been little CAL research on T-unit lengths and ratios. Jin attempted to find out if T-unit measures were useful for describing Chinese, but found no significant difference across a wide range of NNS proficiency levels or native speakers. Jin posited that T-unit measures are not good indicators of language development in Chinese since it is a topic-prominent language. Yuan (2009, p.126) confirmed these findings, as T-unit length and T-unit ratios were actually less complex for NS than NNS speech.

Jin (2007, p.37) proposed a new language division, the **terminal topic-comment unit (TTCU)**. A TTCU is a set of clauses with the same topic but for which the topic appears only in the first clause, with subsequent clauses manifesting the topic in a chained fashion via zero-elements.<sup>12</sup> Jin calculated length-related and density-related syntactic measures using TTCUs and found that covert conjunctions, clauses per TTCU (C/TTCU), and mean length of TTCU (MLTC) showed more complexity at high proficiency levels, but via other scales (e.g. T-unit length) the same language could appear less complex.

**Syntactic variety:** Yuan (2009) noted that “compared with fluency and accuracy measures, complexity measures that have been widely employed in L2 English research are more difficult to apply directly in L2 Chinese, especially the measures at the syntactic level,” and that “so far, no researchers have used syntactic variety to measure L2 Chinese” (p.127). Studies of English syntactic variety often look at verb form variety (e.g. tense, modality, voice), but such inflection-dependent indices are not helpful when analyzing Chinese. Yuan suggested that one way of looking at syntactic variety in Chinese would be developing a measure of known grammatical patterns, such that speech with syntactic variety would

---

<sup>12</sup> For a detailed definition and discussion of zero elements and covert conjunctions, please see Jin (2007).

have a greater density of different grammatical patterns or choices, but acknowledged the difficulty in operationalizing such a measure.

### 2.3.2 CAF measures in the CAL context: Accuracy

Table 2-3 lists the main accuracy measures suggested or investigated in the field of CAL studies, along with variations on the operationalization parameters or interpretation of those measures.

**Table 2-3: Accuracy measures used in recent CAL studies**

Index / construct / measurement	Variation	Notes on calculation	Recent studies		
			Yuan 2009	Yuan 2010	Guo 2007
Error frequency	<i>Error-free clause ratio, aka percentage of error-free clauses</i>	(error-free clauses) / (total clauses)	✓	✓	✓
Error frequency	<i>Error density</i>	(number of errors) / (length of production)	✓		✓
Self-repairs		total number of self-repairs	✓		

According to Yuan (2009), “only few studies in L2 Chinese have employed global accuracy measures” (p.121); in the same study, the author asserts that “the vital challenge...is to find a way to operationalize the measures, namely, what should be counted as an error and what should not be” (p.120). Some reasons for this difficulty in operationalizing the construct in a consistent way are the need for rater agreement to ensure that the operationalized parameters are well-elaborated and repeatable, and the time-consuming, manual nature of global accuracy measures (looking at all errors in a text requires carefully combing through the entire text, from top to bottom). Typological issues arising in the Chinese context include the fact the inflection-based accuracy indices such as those used in Indo-

European languages (e.g. verb tense) are not suitable for Chinese, as discussed above in the context of syntactic variety measures (see sections 2.2.4.1 and 2.3.1). Word order tests are also not directly transferable, as Chinese word order is considered flexible in many situations, particularly in oral speech where emphasis and topic-prominence can lend context to otherwise ambiguous or non-standard word patterns.

Details of the methods of measuring accuracy depicted in Table 2-3 are discussed in greater detail below.

**Error frequency:** Yuan (2009) showed how NS and NNS speech compared in terms of percentage of error-free clauses (number of error-free clauses / total number of clauses) and error density (an error per word ratio; Yuan actually used total words / number of errors). The NS data was determined to have no errors, and it was provided as a contrast to the operationalization of the NNS accuracy measure.<sup>13</sup> In Yuan's (2010) study, the sole global accuracy measure chosen was error-free clauses divided by 'means of clauses', but the 'means of clauses' parameter was not explained (possibly it refers to the mean number of clauses measured per subject in each group). That study found no significant difference in accuracy among groups with different task requirements. Yuan's (2009) paper also mentioned Shi's (2002) observations of a single subject whose error frequency dropped over a 7-month period, and Huang and Qian's (2003) use of error density (number of errors / standard duration or length of production), though Yuan noted the complication of operationalizing error identification in both studies.

Guo (2007) looked at two measures that fall into this category: percentage of correctly pronounced syllables (encompassing initial, final, and tone), and ratio of error-free T-units (REFT, based on

---

<sup>13</sup> Given that the NS sample was a transcription of informal oral language, it is unclear what standards were used to determine grammatical accuracy beyond judgments by two NS raters, and systematic criteria for these judgments, if used, was not cited (indeed, it was not specified whether or not the NS raters examined the NS speech at all).

lexical and grammatical correctness), although it was not clear how the results were interpreted as the raw measures were not included in the paper. However, Guo focused on these measures only as indicators of fluency, and judged that no results other than time-related indices were worthwhile employing for characterizing fluency (see section 2.3.3 for discussion of the time-related fluency indices from this study).

**Self-repairs:** Yuan (2009) illustrated a simple comparison of self-repair measures for NS and NNS speech; in the short samples contrasted, the frequency of self-repair and fillers in the NNS sample was 7 times higher than in the NS sample. Yuan did not attempt to interpret the results, and indeed elsewhere (p.119) further muddled the waters by suggesting the self-repair measure gives a “positive” idea of the degree of a speaker’s self-awareness (i.e. how much attention to form a speaker invests) or conservativeness (i.e. if a learner chooses to use easier formulations; though not fully elucidated, Yuan implies that a more conservative learner would exhibit fewer self-repairs and a lower score in the complexity dimension as he or she would have been more careful in production). Yuan also mentioned that Shi’s (2002) study noted self-repair numbers followed a U-shaped curve for the subject of that study over a 7-month period. Guo (2007) also mentioned this measure but deemed it unhelpful for assessing fluency, the focus of that study.

**Error type and error gravity:** Although Yuan (2009) mentions that previous studies by Shi (2002) and Huang and Qian (2003) have attempted to categorize error types, they are deemed to be difficult to operationalize. Shi (2002) examined correct use of 22 essential sentence types (such as rhetorical questions, ‘be’ sentences with 是, ‘have’ sentences with 有, etc.); Yuan cautions that further research is needed before such a set of measures might be considered reliable and comprehensive enough to be a useful measure of global accuracy. Huang and Qian 2003 differentiated between lexical and syntactic errors, but Yuan points out that the lexical / syntactic divide in Chinese grammar is a contentious one.

### **2.3.3 CAF measures in the CAL context: Fluency**

Table 2-4 lists the main fluency measures suggested or investigated in the field of CAL studies, along with variations on the operationalization parameters or interpretation of those measures.



**Table 2-4: Fluency measures used in recent CAL studies (continued on page 33)**

Index / construct / measurement	Variation	Notes on calculation	Recent studies		
			Yuan 2009	Yuan 2010	Guo 2007
Speech rate	<i>Speech rate</i>	syllables / second	✓	✓	✓
Speech rate	<i>Speech rate while talking, a.k.a. 'rate of speech' / articulation rate</i>	(number of non-repeated syllables) / (speech time - length of pauses)			✓
Speech rate	<i>Quality production rate or Mehnert production rate</i>	(total words - replaced, repaired, reformulated words) / time	✓		
Pauses	<i>Total false pauses</i>	(total number of false pauses)	✓		
Pauses	<i>Mean false pause length</i>	(total length of false pauses) / (number of false pauses)	✓		✓
Pauses	<i>Phonation/time ratio (PTR)</i>	(Total time - total silent pause time) / (Total time)			✓
Pauses	<i>Mean Length of Runs (MLR)</i>	(Total syllables of speech) / (Total false pauses)			✓
Pauses	<i>Mean length of utterance</i>	(Total time - total false pause time) / (Number of false pauses)			✓
Self-repairs	<i>Self-repairs</i>	total number of self-repairs	✓		
Self-repairs	<i>Repairs per 100 syllables (R100)</i>	(total number of self-repairs) / (100 syllables)			✓

Index / construct / measurement	Variation	Notes on calculation	Recent studies		
			Yuan 2009	Yuan 2010	Guo 2007
Self-repairs	<i>Ratio of pruned length to total length (RPL)</i>	(Total 'pruned' syllables (repair syllables)) / (total syllables produced)			✓
Dysfluency	<i>Disfluency</i>	(total number of self-repairs, fillers, and English words) / (total words <i>or</i> total time)	✓		
Total words				✓	

*(continuation of Table 2-4 from page 32)*

According to Yuan (2009, p.118), “not many studies on L2 Chinese have used fluency measures to evaluate learners’ performance;” in further research Yuan cautions that decisions about standards (e.g. words per minute or syllables per second) and recognition of word boundary issues in Chinese are important to keep in mind. Other theoretical issues include Guo (2007) noting that pauses can be attributed to linguistic factors (i.e. ‘how do I say this?’) and cognitive factors (i.e. ‘what should I say?’), but that fluency measures by necessity conflate these (that is, it can be difficult to know what’s really happening in the speaker’s head, even perhaps for the speaker). Another potential issue is that when manually coding disfluencies, a relatively small change in coding counts can introduce a large effect for small text sizes, so cross-study comparisons might be difficult without stringent protocols for coding.

Details of the methods of measuring complexity depicted in Table 2-4 are discussed in greater detail below.

**Speech fluency:** Yuan (2009) demonstrated calculation of raw speech rate and quality production rate (excluding “replaced, repaired, or reformulated” words, p.114) in units of syllables per second for NS and NNS speech, and unsurprisingly concluded that NS speech is faster, particularly when self-repairs were taken into account (the NS example exhibited no self-repairs, thus the quality production rate was equal to the speech rate). Yuan mentioned that Zhang (2001) examined fluency as a function of anxiety level in Chinese speakers using a speech rate measure to demonstrate that high anxiety levels led to slower speech. Yuan’s (2010) study of written production attempted to measure fluency by looking at the number of words produced per minute, finding that it dropped as expected for the focus-on-form experimental group.<sup>14</sup>

Guo (2007) found that of the indices studied in that paper, those related to time were very good indications of fluency, but all other measures attempted were not. Guo’s time-based speech fluency measures were

---

<sup>14</sup> Yuan (2010) acknowledged that “word boundary is not always clear in Chinese” and explained that word count was determined in that study by “strictly following the HSK guidelines when segmenting ... sentences into words” (p. 77).

speech rate (in that study called 'speaking rate'), 'phonation/time ratio' ( $((total\ time - total\ pause\ time) / total\ time)$ ), and 'articulation rate' ( $(total\ syllables / (total\ time - total\ pause\ time))$ ), the latter two of which might also be considered breakdown fluency measures.

**Breakdown fluency:** There are a myriad of ways to operationalize this dimension of fluency. Yuan (2009) and Guo (2007) measured mean length of false pauses, where the threshold was 0.5 seconds and 0.3 seconds, respectively. Yuan (2009) noted Zhang's (2001) use of number of false pauses within 100 syllables, and measures for number and mean length of false pauses within clauses and between clauses. Yuan (2009) also explained that Shi (2002) calculated total false pauses, finding that the subject of that investigation showed a marked drop in false pause rate (per clause) over the course of the study.

Again, based on Guo's (2007) observations on time-based measures, (see 'speech fluency' section, above) Guo found only mean false pause length to be a useful indication of fluency.

**Repair fluency:** Yuan (2009) mentioned that Shi (2002) noted self-repairs for the subject of that study, but did not quantify the results. Yuan (2009) operationalized repair fluency as **disfluency**, defined as  $(total\ number\ of\ self-repairs,\ fillers,\ and\ English\ words) / (total\ words)$ . This makes disfluency a sort of companion to **quality speech rate**, described above in the section on 'speech fluency'. The two measures Guo (2007) calculated to represent repair fluency did not merit further research in that author's opinion.

## 2.4 Gaps in the current research and directions for future research

Though research on complexity, accuracy, and fluency in the CAL context is not new, it is still an undeveloped field. Unlike measures for English and other European languages with longer traditions of SLA scholarship, there does not yet seem to be a consensus on which measures work to classify NS and NNS Chinese proficiency. Future research is needed in a number of areas; the following areas are all addressed to some extent by this thesis:

- Applied linguistics research on automating calculation of CAF measurements and presenting them in a useful format for researchers and educators;
- More qualitative investigations similar to Yuan (2009) so that researchers, learners and teachers understand how the operationalization of these parameters maps onto the actual language use;
- More studies on aligning CAF measures with objective measures such as HSK, TOCFL, ACTFL OPI, and CEFR;
- More NNS studies and larger NNS data sets to verify or challenge the results found in the few studies that have been done; and,
- Tests of the advances proposed in CAF studies on other languages to see if those improvements work in the Chinese CAF context (e.g. D value, MATTR).

Further research in the following areas, outside the scope of this thesis, would also be helpful:

- Explorations of novel, Chinese-specific CAF indices, such as Jin's (2007) TTCU measures, or tonal accuracy measures;
- Investigation of the link between local or specific indices (e.g. discrete sets of grammatical checks that are easy to code) and global accuracy indices (which are difficult to code efficiently for large data sets);
- Studies of NS language to provide baseline data for L2 acquisition research and a platform for comparing NS and NNS data;
- More comprehensive work on how task conditions affect CAF in the CAL context, as there is a substantial body of research on this in other language contexts for comparison;
- Incorporation of research done in other areas of CAL acquisition, such as error analysis, order of acquisition research, and discourse competence research.

As already mentioned, this thesis attempts to address some of these issues, and for others it discusses possible next steps for other research (see Chapter 5).

## **2.5 Summary**

This chapter introduced the major assessment types used in measuring oral language proficiency, with a focus on the CAL context. It explained the concept of complexity, accuracy, and fluency (CAF) analysis, and explored in detail CAF research in the field of teaching and learning Chinese as an additional language. Many of the techniques described herein will be revisited in Chapters 3, 4, and 5; the foundational research discussed in this chapter has given a basis from which to test the value of specific parameters of operationalization, explore new techniques, and posit on how existing techniques can be improved through automation or calibration research.

## **Chapter 3. Data and methodology**

### **3.1 Introduction and context of data collection**

The analysis in this thesis is based on oral interviews, self-assessments, and reported linguistic data of five adult English-speaking CAL learners collected as part of a larger project (Duff et al., forthcoming) investigating their experiences and trajectories. The larger project included learning narratives, cooperative analyses of narrative themes, metadata analysis of the collection and discussion process, and further descriptive data collection to give a wide and at times also detailed portrait of the five participants. The oral interviews and self-assessments were collected to characterize learner proficiency, and some of the work in this thesis was included in the manuscript of that monograph in brief; the analysis presented here is a more thorough look at the work required to characterize the participants' proficiency based on the oral data and cross-reference it with reports of standardized test scores achieved in the past along with self-assessments. Where parts of this analysis were performed by co-authors of the monograph they are credited for their work in this thesis; some of the concepts of analysis and methodology were discussed as a group and input was invited from everyone, but unless otherwise stated the choice of methods, appropriate parameters, scoring, computer programming, graphing, and analysis were prepared by me.

The following are the research questions which guided the work, reproduced from Chapter 1:

1. What techniques and tools are efficient and useful for characterizing linguistic patterns, sophistication, and variability in oral production data in Chinese as an Additional Language (CAL)?
2. Do these tools help to characterize meaningful differences in Chinese oral production and knowledge among learners:
  - representing a range of proficiency levels and previous formal language study?

- in capturing differences over time, i.e. language development?
  - in comparing against benchmarks, or otherwise objectively determining proficiency?
3. What kinds of analysis can best be facilitated by computer tools such as those developed for the field of computational linguistics? What kinds of analysis might benefit from further developments in computer-assisted tools? What aspects (lexis/morphology, syntax, fluency, discourse) are not easily measured via computer-assisted tools, and could be investigated further in future research? What features need to be improved in the realm of freely available computational linguistic software for the purposes of describing oral language use and proficiency?
  4. What are some of the pedagogical implications of the different tools and analyses available for the linguistic analysis of oral Chinese? In what ways can such analyses assist Chinese language instructors and learners, including those learning abroad in Chinese-speaking regions? What implications are there for textbook design and course delivery?
  5. How are attempts to analyze and characterize the learning of oral Chinese different from discussions of oral proficiency development in other languages?

### **3.2 Description of participants**

The five participants were adult L1 English speakers. Four (Ella, Elliott, Roma, and Tim) were graduate students in the Faculty of Arts or the Faculty of Education at the University of British Columbia (UBC), and the other (Patsy) was a professor in the latter faculty. All had multiple years of experience learning Mandarin Chinese and at times living in Chinese-speaking communities. The participants became involved in the study as volunteers and co-researchers as they were all colleagues in the UBC Centre for Research in Chinese Language and Literacy Education (CRCLLE) and wished to collaborate on an investigation of the common threads in the



experiences of CAL learners from a North American English-speaking context. Table 3-1 shows the participants' Chinese learning backgrounds relative to one another.

**Table 3-1: Participants' Chinese learning backgrounds**

	<b>Ella</b>	<b>Elliott</b>	<b>Patsy</b>	<b>Roma</b>	<b>Tim</b>
<b>Primary site(s) of study</b>	China, Canada	Taiwan	China, Canada	Singapore, Canada, China	Taiwan
<b>Details of study</b>	<ul style="list-style-type: none"> <li>• Mix of informal and formal studies</li> <li>• 6 months' studies in Taiwan between 2009 and 2010</li> </ul>	<ul style="list-style-type: none"> <li>• 4 years of formal studies</li> </ul>	<ul style="list-style-type: none"> <li>• Mostly informal studies</li> </ul>	<ul style="list-style-type: none"> <li>• Formal studies included high school, college, and study abroad</li> </ul>	<ul style="list-style-type: none"> <li>• Studies focused on oral language</li> <li>• 6 months in Taiwan studying literacy between 2009 and 2010</li> </ul>

Though not a participant in the study, the sixth member of the CRCLLE research group guiding the study and preparing the monograph (Rachel) was an applied linguistics researcher and the CRCLLE coordinator during the two years that the data was being collected and analyzed for the monograph project. Rachel has a background in Chinese language pedagogy and her first language is Chinese.

All participants provided their informed consent via the procedures established by The University of British Columbia Office of Research Services Behavioural Research Ethics Board to include their real names in this thesis, as per their preference.

### **3.3 Data used in this thesis**

#### **3.3.1 Oral interview data**

The principal data collected for CAF analysis for this thesis and for the monograph were one-on-one, informal, semi-structured oral interviews. The interviews were loosely designed by the group to elicit interesting experiences and a variety of language in a comfortable, low-stakes environment, with friendly scaffolding from the interviewer, Rachel.

Two interviews were conducted per participant: one in the (North American) summer of 2009 and one in the summer of 2010. It was hoped that the intervening year might bring about observable differences in participants' speech, as between the two interviews some learners pursued further Chinese studies, while others did not engage any differently with the language after one year. The format of the interviews is described in Table 3-2.

<b>Oral interview format in 2009 and 2010</b>
1. "Warm up" chat
2. Question & Answer: personal questions about the participant's life and Chinese learning experiences
3. Picture description: given a brief oral prompt, discuss a picture in the form of a short monologue. (In 2009 there was one picture; in 2010 participants were asked to discuss two different pictures separately.)
4. Describe a cultural work: Participant asked to prepare for an extemporaneous description of a book, movie, or television show of his or her choice
5. "Cool down" debrief

**Table 3-2: Oral interview format in 2009 and 2010**

The interviews, typically around 45 minutes in length, were recorded using a digital audio recorder. The 2009 interviews were transcribed by Rachel; the 2010 transcripts were first drafted by another UBC graduate student from the Department of Asian Studies, then checked by Rachel so they conformed as much as possible with the 2009 conventions she had used. Rachel also found that due to the 2010 transcriber's unfamiliarity with the participants' speech mannerisms and the interview topics and contexts there were certain utterances that required alterations or clarifications, so the transcript used in this study is the one she edited. The transcriptions were made using a personal computer, via the operating system's default Chinese Input Method Editor.

These transcriptions were the basis of most of the CAF analysis, using only the data from questions 2, 3, and 4; therefore the data for analysis were only a portion of the total interview length, and in the case of fluency a very short portion (see below). Specifically, the warm-up and cool-down sections were not analyzed, as they were designed to ease the participant into speaking Chinese (they all operated in daily life primarily in

English, in an English environment), establish rapport with the interviewer, and in the case of the cool-down section, elicit retrospective commentary in English as well as Chinese about the interview itself and the participant's performance.

For fluency analysis, the audio files were used in concert with the transcriptions. During the original preparation for the CAL learner monograph the group chose to use guidelines from the 'study abroad' sub-field of SLA research as guidelines on how to design and analyze measurements of fluency. Rachel, Dr. Duff, and I discussed which methods to use, and Rachel finalized the parameters and performed the analysis based on modifications of the measurements and methodologies described in Collentine & Freed (2004), Freed (1995), and Yuan (2009), with some logistical and measurement assistance from me. The suitability of the measurements chosen and their relationship to the CAF resources discussed in section 2.3 will be further explored in Chapters 4 and 5.

Rachel chose two-minute extracts of interview sections 3 and 4 (see Table 3-2) from each interview, for a total of twenty separate two-minute-long speech samples (five participants, two questions per interview, two interviews per participant, namely 2009 and 2010). The two-minute criterion was chosen in order to obtain a considerable amount of speech, but was limited by the amount of data available, as some participants did not speak for much longer on any given task; this criterion resulted in speech samples ranging from 379 to 630 Chinese syllables in length. Rachel and I then made measurements of speech fluency and calculated some fluency indices based on these measurements (see section 4.4). We used Audacity<sup>15</sup> to time the two-minute extracts and manually calculate the pause measurements.

Only one researcher (Rachel) took fluency measurements from the interview data (i.e., there was no inter-rater check, although Audacity is quite precise in e.g. pause length measures). This analysis was considered exploratory and we did not attempt to make statistical inferences in the monograph; we merely intended to

---

<sup>15</sup> Audacity is free, open-source, user-friendly audio software with a rich feature set: <http://audacity.sourceforge.net>.

provide this analysis as a way to engender interest in the methodology from other researchers, and to consider its methods for this project.

CAF analysis was facilitated by transforming the transcripts via various automated processes (e.g. removing punctuation, isolating learner speech, and parsing the raw character stream into words). The bulk of the automation in text manipulation and analysis was performed using computer scripts in the Perl programming language,<sup>16</sup> with some follow-up sorting, counting, and calculations performed in Microsoft Excel. The general details of these processes are described herein, with examples given in some cases to illustrate:

1. Transcripts were reduced to the substantial sections 2, 3, and 4; sections 1 and 5 (warm-up and cool-down) were removed.
2. Transcripts were stripped of interviewer speech, and transcription comments (which were included in parentheses to indicate elements such as laughter, pausing, transcriber assumptions, and so forth).

This step was performed mainly using Perl-language computer scripts, with manual tweaking on some files to ensure the scripts operated consistently across the entire sample set.

3. The remaining text, representing solely substantial participant speech, was segmented using ICTCLAS 5.0.<sup>17</sup> The segmenter tokenized the text; that is, the raw stream of Chinese characters was broken into tokens, in this case words, as accurately and consistently as possible by ICTCLAS.<sup>18</sup> For example, the

---

<sup>16</sup> The Perl programming language is a scripting language that is widely used for text processing and rapid prototype development: <http://perl.org>.

<sup>17</sup> ICTCLAS (Institute of Computing Technology Chinese Lexical Analysis System) is a freely available Chinese segmenting library developed by the Institute of Computing Technology, Chinese Academy of Sciences: <http://ictclas.org>.

<sup>18</sup> ICTCLAS and other similar Chinese segmenting software attempt to break up a stream of Chinese characters into words and identify each word's part of speech. However, the concept of 'word' in Chinese is a contested one, and strictly speaking such segmenting tools break texts into 'tokens' or 'n-grams' (groups of  $n$  syllables/morphemes), that is, clusters of morphemes that when taken together represent a single, indivisible idea. Whether or not each token is also a word is a matter of debate, and may depend on the frame of reference (e.g. is the token exactly equal to a Chinese reference dictionary entry? Is the token a compound of known dictionary entries?). For further discussion of what constitutes a word in Chinese and how segmenters deal with these issues, see for example Chen & Liu (1992), Packard (2000), Huang et al. (2008), and Lazarinis et al. (2009).

sentence “我目前在学习的是...Berkeley 大学所提供的...um...” (‘The [one] I’m currently studying is... [the one from] Berkeley... um...’) becomes “我 目前 在 学习 的 是 ... Berkley 大 学 所 提 供 的 ...?” (i.e. constituent words / tokens are now separated by spaces for analysis).<sup>19</sup>

4. Segmented files were then stripped of English words (which were present on account of occasional code-switching behaviour), and any non-Chinese characters in the written transcript (principally consisting of punctuation) so that the following steps gave accurate word counts. (These words and punctuation were useful to the segmentation software in step 3, but complicate word type analysis, frequency analysis, and so forth.) For example, the phrase mentioned in step 3 becomes “我 目前 在 学习 的 是 大 学 所 提 供 的.”
5. Segmented, cleaned files were analyzed for complexity measures using Perl scripts and algorithms designed to calculate syllable, token, and unique token counts, and to compare the data with reference materials. Details of these measures are described in Chapter 4.

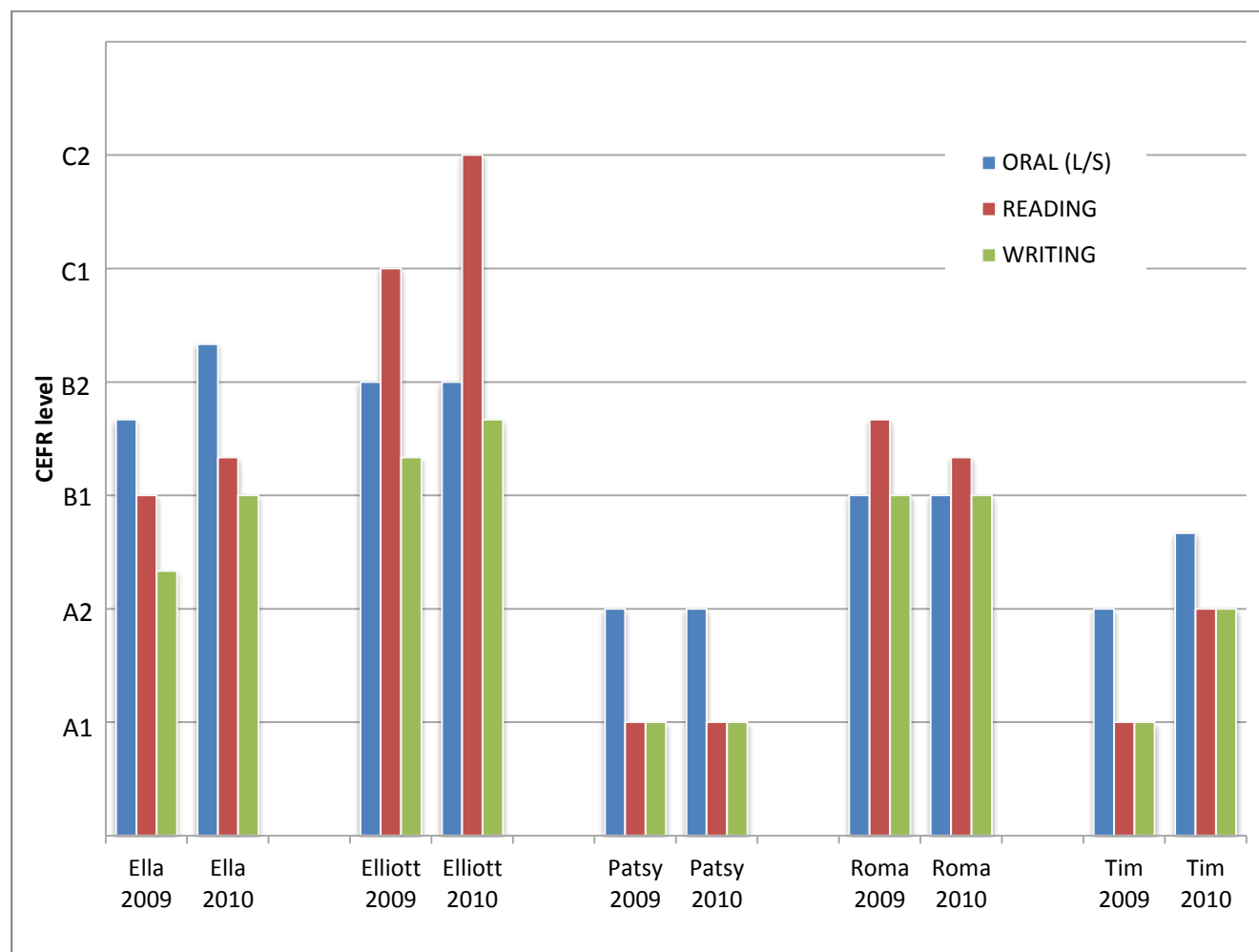
### 3.3.2 Self-assessment data

The group also deemed it vital to characterize proficiency in alternative ways, in order to determine what different methods of assessment highlighted or captured in terms of describing CAL proficiency. Rather than have participants take a standardized test such as the HSK or TOCFL, the group decided that self-assessment using the well-known and robust CEFR would be most appropriate, as literacy varied from virtually nil to advanced, rendering a standardized test impracticable as a group. The CEFR, on the other hand, was a well-understood, tested, and aligned framework (see section 2.2.1 for a description of the CEFR), particularly for European languages, with detailed tools that one could use to self-assess based on competencies and concrete examples of proficiency. Each study participant self-assessed various sub-skills using established self-

---

<sup>19</sup> Note that in this example, ‘...’ represents a pause in speech, not ellipsis.

assessment grids (see Table A-2 and Table A-3 in Appendix A for more details), and used a further refinement of + (a plus sign) to indicate strong confidence in meeting the descriptors for a given level or – (a minus sign) to indicate low confidence in meeting the descriptors for a given level. The results of the CEFR self-assessments are shown in Figure 3-1, with +/- indicated as a minor step above/below a horizontal grid line, respectively.



**Figure 3-1: CEFR self-assessments for each participant in 2009 and 2010<sup>20</sup>**

<sup>20</sup> Participants are listed in alphabetical order in all lists, tables, and figures in this thesis.

Examining Figure 3-1 allows us to get a sense of the participants' variation both across skills areas and over time, and see how the events of their lives during the time of the study correlated with their Chinese proficiencies.

- Ella, Elliott, and Tim judged their proficiency as improving in some respects over the course of a year, whereas Patsy and Roma did not; given that Ella and Tim pursued full-time study in Taiwan over the period between the two assessments, the fact that they reported an increase in all three sub-skills makes sense, whereas Elliott's continued love of reading may account for the increase in his reading sub-skill only.
- In contrast, Patsy and Roma, who were not engaged in formal study, did not report much change in their abilities from 2009 to 2010.
- Ella, Patsy, and Tim believed that their oral skills were more developed than their literacy skills, while Elliott and Roma felt their reading proficiency was higher than their oral and writing proficiencies.
- By 2010, Patsy's assessment of her literacy skills was lowest among the participants, and Elliott's the highest; Patsy's assessment of her oral proficiency was also lowest, while Ella's was the highest.

For the purposes of this thesis, the important sub-skill under examination is oral proficiency. It was used as the baseline to which CAF analysis could be compared to determine whether or not different types of analysis were useful or potentially misleading.

### **3.3.3 Standardized tests**

To triangulate data in the current study, some of the participants reported results from past standardized tests. These test results are summarized in Table 3-3. For a brief description of the tests, please refer to section 2.2.1. These standardized test scores were then compared with the CEFR self-assessments described in section 3.3.2 to validate the adequacy of the self-assessments as benchmark data for the CAF analysis.

**Table 3-3: Participants' results on HSK, TOP, and TOCFL standardized tests.**

Participant	HSK (old)	TOP (old)	TOCFL (current)
Ella	Intermediate (7) 2006	Intermediate (3) 2009	
Elliott		Advanced (6) 2007	
Patsy	Elementary (5) 2008		
Roma			
Tim			Beginner 2010

*Note: Entries show the level name, (sub-)level number (if applicable), and date passed.*

Note that standardized tests do not characterize in fine detail if the learner is not literate in *Hanzi* (Chinese characters), or at least not before the advent of pinyin versions of the tests (though test-takers still require literacy skills in pinyin). This explains why neither Patsy nor Tim had taken any intermediate or advanced tests, as they did not learn literacy skills in a systematic way (though Tim began to learn by the end of the study, and did take the 2010 TOCFL), and in Patsy's case oral components were not available at the time she was principally learning Chinese.

We can see that though the tests were taken at different times, they appear to support the self-assessments summarized in Figure 3-1, above. We can examine these in light of the calibrations summarized in Table 2-1 (see section 2.2.1).

- Elliott's TOP score, corresponding to C1, does appear to match his self-assessment of a C1 in reading, though he believed his oral and written skills were in the upper-B range. This suggests either underestimation of abilities or attrition following the 2007 test and prior to the interviews in 2009 and 2010.
- Tim's 2010 TOCFL result, calibrated to A2, matches his self-assessment of A2 proficiency in reading and writing; he felt more confident in his oral proficiency, and indeed since standardized tests require the learner to choose one level to test at, his TOCFL score may in fact not fully represent his abilities; he took the Beginner test as he was focusing on going from essentially zero literacy skills to an A2 (CEFR) level at the time, and thus the test can be seen to reflect that proficiency more than anything. He had



been using his oral abilities for years prior to these interviews, so it makes sense that he would assess them as more developed than his literacy skills.

- Roma's 2008 test results can be calibrated in the B2+ level, which makes her self-assessment, clustering around B1 but reaching as high as B2-, seem low; seeing as she took her HSK test while living in China, it is possible that, similar to Elliott's case, she either underestimated her abilities at the time of the interview, or experienced attrition in the time between the standardized test and the current study.
- Ella's Intermediate HSK results suggest a C1 level of proficiency, higher than the average B1 score she reported in her 2009 self-assessment; this may suggest lack of confidence or attrition, as in Elliott's and Roma's cases outlined above; note also that her HSK results were three years old by the time of the current study, the largest gap for any participant. Following the 2009 self-assessment Ella moved to Taiwan and studied the traditional orthography, thus her B2-calibrated results for the TOP make sense in light of the time required to learn the new orthography; they also correlate better with her 2010 self-assessment, which lends support to the idea that the self-assessment is a valid reference point for her proficiency.
- Finally, Patsy did not report any standardized test results, unsurprising in light of her minimal experience learning and using Chinese character-based literacy, and the lack of an oral proficiency subcomponent on the earlier HSK test.

Given the standardized test scores appear to reinforce the self-assessments provided by the participants, they lend credence to the adequacy of the self-assessments in representing a baseline that the CAF analysis can be compared with. One interesting point to note is that in all cases where the previous exam results were compared with current self-assessments the exam results placed the participants' proficiency one level higher (using the CEFR global scale) than their self-assessment. In the two cases where tests were taken within a year of the second self-assessment (Ella in 2009, and Tim in 2010), the results of the tests (as aligned to the CEFR, see Table 2-1) accord with the self-assessed scores. This might indicate that Ella's, Roma's, and

Elliott's 2009 self-assessments were conservative, or that attrition was in evidence; it might also have indicated that the tests were not calibrated well to the CEFR, but without a larger data set of test-takers and larger CEFR alignment studies, it is hard to make any conclusions from these observations.<sup>21</sup>

### 3.4 Analysis

The interview data was analyzed for indicators of linguistic proficiency and individual linguistic development. Indicators chosen were those suggested by previous scholarship and any that lent themselves to Chinese analysis that had not been used in the past. The study mainly focused on lexical, syntactic, and fluency phenomena in the data, with the following purposes guiding the methodology:

1. **Where possible, devise automated ways of accomplishing tasks otherwise tabulated, coded, or calculated by hand.** This involved using computer software such as word processors to search through large amounts of text in a fast and accurate manner, creating scripts to perform repetitive or complex calculations on large sets of data, and employing specialized computational linguistic software libraries where available to leverage tested, state-of-the art, standardized methods of analyzing speech or text. The painstaking, manual nature involved in parsing words and clauses, manually tallying syllables, self-repairs, English words, and so on, as described by Yuan (2009), by necessity makes research expensive and time-consuming and dissuades educators from attempting such analyses. Although this study can only take one small step towards helpful shortcuts, it aims to bring the field at least that one step closer to widespread understanding and rapid application of these analytical techniques, in the hopes that ever larger data sets can be processed more consistently and quickly in the future, and by a wider variety of investigators.

---

<sup>21</sup> Note as well that standardized tests tend to focus on grammatical, lexical, and character knowledge, while the CEFR focuses on broad functional capabilities.

2. **Focus on *presence* of a pattern or phenomenon in keeping with the ‘can do’ spirit of the CEFR self-assessments** (whereby learners report on their capabilities in a language rather than their limitations) and the larger CAL study which this thesis informed. This mainly meant avoiding characterizing errors or absences, where possible, and attempting to concentrate on positive indicators of proficiency and performance. Some analysis methodologies did require accuracy checks, so sometimes a particular instance was discounted if not considered an acceptable use of the item in question (for instance, when analyzing polysemous morphemes for the analysis described in section 4.2.2, if a given occurrence of the morpheme was used inaccurately, or in an ambiguous way, it was discounted from the analysis; e.g. use of the completion/change morpheme 了 *le* in an ungrammatical context). For lexical usage, however, trust was placed in the accuracy of the computational linguistic programs used to parse or segment the speech, and the assumption was made that all instances of a particular lexical item were contextually appropriate (native-like). Another way to justify this position is the assumption, for the purposes of this project, that at least exhibiting knowledge of and desire to use a structure or word is an attestation to some level of proficiency involving that structure or word. (An additional advantage of investigating only positive use of structures or words is that finding them and examining their context is much easier than pinpointing errors, especially the *absence* of language required for native-like communication of a thought.)
3. **Be flexible in the use of a given unit of measure, but clear in the methodology of calculating an index.** Although important to try to replicate as much as possible conditions under which past and future studies might be directly compared to results obtained in the current study, it was acknowledged that very little work has been done in the application of CAF analysis to CAL. It was considered more important to explore as many promising methods as possible, and crucially to describe them well and provide a clear rationale for more nuanced and better-informed follow-up studies, than to blindly calculate the exact figures used in other studies or narrow the data range in any way to accommodate fac-

ile comparisons. One example is the decision to segment speech into syllables in some cases (e.g. when examining fluency, so that speech rate was less dependent on mean word length), and words in others (e.g. when examining complexity, as syllables or morphemes without their immediate context collapse a range of meanings and sophisticated uses into a single symbol when transcribed, and so do not lend themselves to complexity analysis).

### **3.5 Summary**

This chapter presented the greater context of the data examined in this study. It explained what and how data was collected and transformed for analysis, and provided a general overview of the methodological principles brought to bear to determine which CAF measures to investigate for this study. Given the guiding principles described in section 3.4, some of the CAF analysis types enumerated in Chapter 2 were deemed to be too time-consuming, too difficult, or otherwise outside the scope of this thesis. The specific indices, methods, and parameters that were devised for CAF analysis of the CAL speech data used in this thesis are detailed in full in Chapter 4.

## Chapter 4. CAF analysis – results

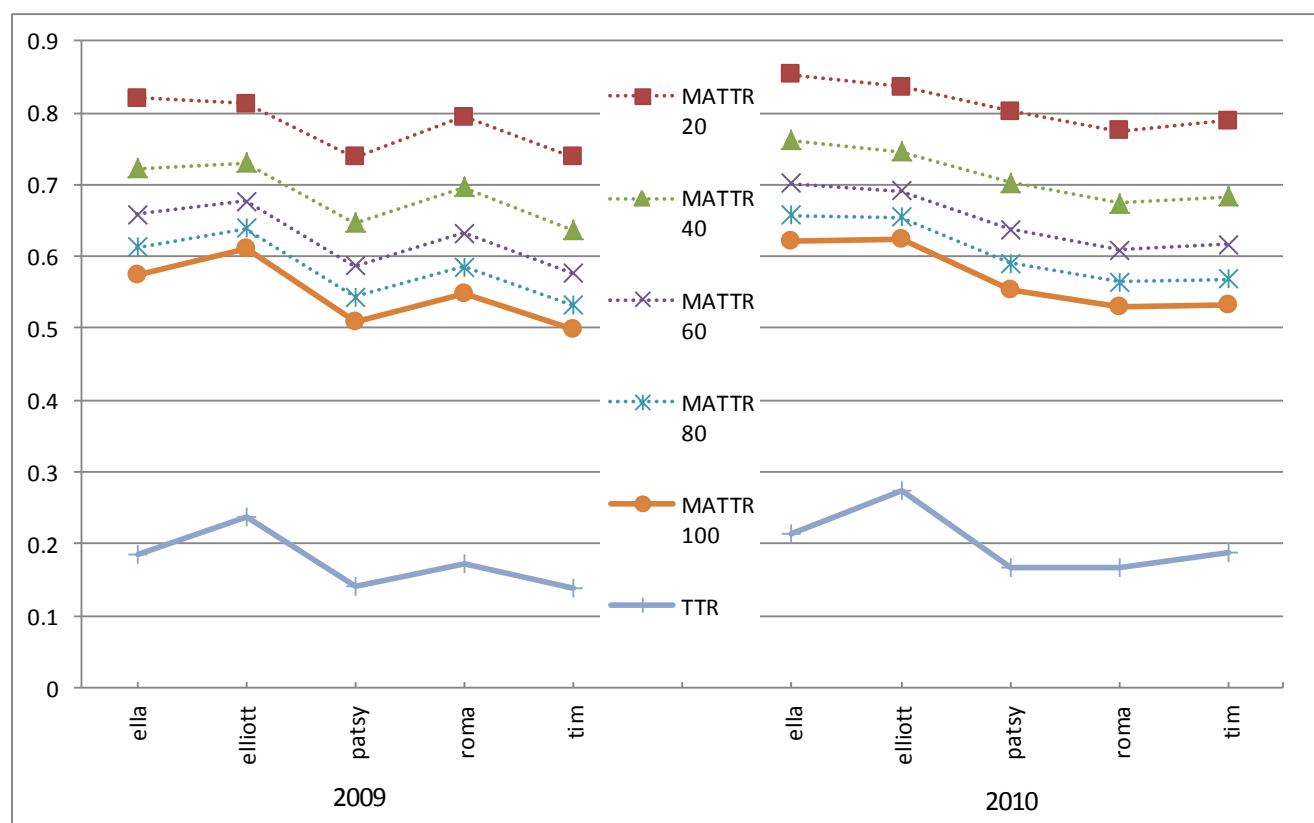
### 4.1 Introduction

This chapter presents the results of calculating various CAF measures for the interview data introduced in Chapter 3. Where possible it attempts to compare these results with the CEFR self-assessments each participant provided in 2009 and 2010, in order to determine at a glance whether a specific CAF index appears to be indicative of proficiency in the CAL context (and for this data set in particular). Parameters and judgments in operationalization are described in this chapter; Chapter 5 will go into further detail on the implications of the results presented here.

### 4.2 Complexity measures

#### 4.2.1 Lexical variety

I examined two measures of lexical variety, **type-token ratio (TTR)** and **moving average type-token ratio (MATTR)**. TTR is a more straightforward calculation, and easier to mentally picture, being the total unique words (types) divided by the total words (tokens) produced; a higher TTR would indicate greater variety in lexical expression. MATTR attempts to account for the compromising effect of varying text lengths (see 2.2.4.1), and is essentially a set of TTR calculations incorporating a small window of the source text averaged over the entire text – it aims to give an indication of how consistently someone uses a variety of words from thought to thought, topic to topic. For this study I followed the MATTR algorithm proposed by Covington & McFall (2010), tried a number of window sizes to see the effects of the calculation, and compared MATTR to TTR for the interview data in this study. The TTR and MATTR calculations are shown in Figure 4-1.



**Figure 4-1: TTR vs. MATTR calculated for different window sizes of 20, 40, 60, 80, and 100 words. MATTR 100 is highlighted as being closest in discriminatory power to TTR.**

Comparing the data in Figure 4-1 with that of the participants' self-assessments (3.3.2 and Figure 3-1), we can see that Ella and Elliott, who assessed themselves in the B2 and B2+ range for oral proficiency, but not lexical variety specifically, did consistently show the highest scores in the group for TTR and MATTR. Ella's scores rose from 2009 to 2010 consistent with her change in self-assessment (from B2- to B2+), and Elliott's scores rose slightly as well, though the MATTR in particular did not change much.

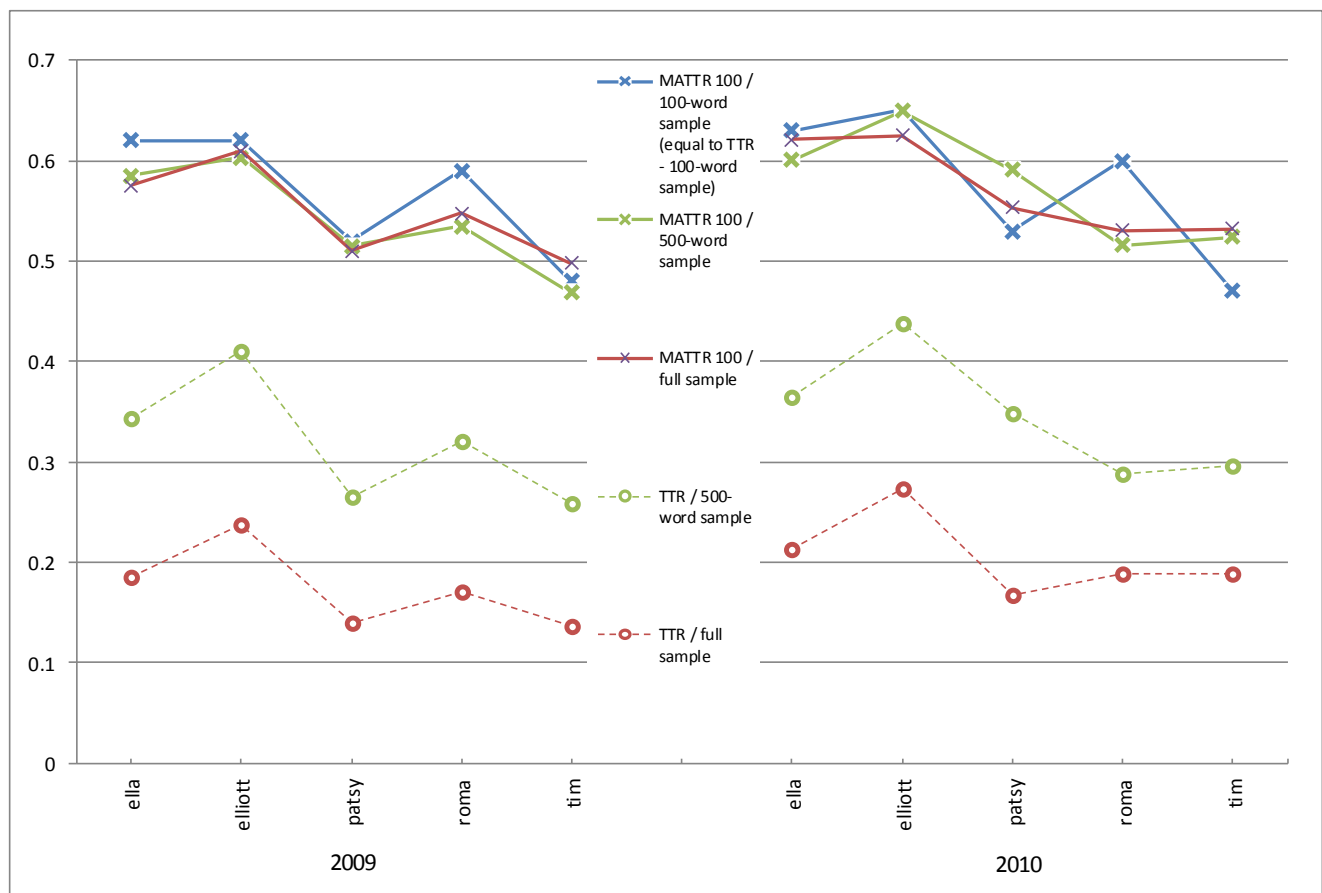
Examining participants in the A2 to B1 range, we see that Tim and Patsy, A2 in 2009, also had similar scores in 2009, and their scores were lower than Roma's who assessed herself at B1. In 2010, Tim's lexical variety score did rise in comparison to 2009, as expected from his higher self-assessment of B1-. Roma's lexical variety was lower in 2010 than in 2009, though not by a large margin; her self-assessment also remained at B1. Patsy's TTR and MATTR scores rose from 2009 to 2010 despite her self-assessment remaining at A2.

Overall, the relative values across participants and as compared to participants' oral CEFR levels do seem to show a consistent relationship, giving weight to the expectation that TTR and MATTR are indications of relative proficiency. If true, one could posit that Patsy's and Elliott's proficiencies did rise somewhat from 2009 to 2010, despite their self-assessments; active participation in Chinese-speaking communities and activities, greater comfort with the interview process, an expanded set of questions in the second interview (see 3.3.1), or more forethought on the topics they were able to prepare for may have been factors. The fact that the CEFR is a global proficiency scale, and is largely based on functional 'can do' criteria which don't have an easy-to-ascertain correlation with lexicon size, may also account for the observed disconnect between CEFR self-assessment and lexical variety: in Patsy's case, the jump from A2 to B1 on the CEFR global scale is quite large, and the changes in TTR and MATTR scores may indicate that her performance improved in a way that her self-assessment on the less-finely tuned CEFR global scale may not capture.

In terms of the appropriate window size for MATTR, Covington & McFall's (2010) discussion is enlightening: a window size of 10 words or less can be used to detect disfluency, 500 is recommended for examining text style, and 10,000 is recommended for determining vocabulary size. I chose to examine MATTR window sizes from 20 to 100, so that there were many non-overlapping windows for the participants' interview texts (the length of each of the ten interviews, once pared down to just participant speech, ranged from approximately 1,800 to 2,500 words) and from the data shown in Figure 4-1 we glean that MATTR 20 is approximately 20% less discriminatory (that is, differentiating between samples) than MATTR 100; MATTR 40 in turn is 10% less discriminatory. MATTR 100 turns out to be nearly identical to TTR in terms of discriminatory range, or the average difference between the high and low values calculated for participants in a given year.

Given its greater discriminatory power, I performed some calculations to see if MATTR 100 was more consistent than TTR for text length. As shown in Figure 4-2, the MATTR with window size 100 did indeed maintain a much more consistent score across samples of different sizes, and would thus be a better indicator for speech samples of varying lengths; the ratio of the change due to text length to the variation in sample scores

was much more pronounced for TTR than MATTR, confirming MATTR shows greater consistency (lower variance) for texts of different lengths. As the word counts for each participant's interviews varied from as low as 1,891 to as high as 2,600, MATTR appears to be a better way to capture cross-sample indices of lexical variety than TTR for the data in this study, and MATTR 100 in particular, considering its discriminatory range as discussed above.



**Figure 4-2: MATTR with window size of 100 words calculated for different sample sizes**

Note that the variance between MATTRs calculated for different text lengths, most observable for Patsy, Roma, and Tim's 2010 MATTR values in Figure 4-2, are most likely due to the choice of samples. The full sample (all words from the interview data) should give the best indication of ability; the other reference samples of 100 and 500 words were created by taking the first 100 and first 500 words, respectively, of the full



sample. The difference in topic choice, fluency, and comfort between the beginning and end of the interviews may account for the variation observed; if samples of 100 or 500 words culled randomly from the full texts were used, that might mitigate some of the variance.

As a token qualitative indication of the value of the MATTR measure, I also calculated the MATTR for a well-spoken native speaker discussing housing issues in China in a multi-party conversational format, as transcribed by a Chinese broadcaster.<sup>22</sup> The native speaker's MATTR 100 value of 0.70 is indeed higher than any of the participants, which matches the theoretical expectation that a lexical variety index should discern between educated native speech and intermediate learner speech, though it must be noted that the NS speech sample was from a nationally broadcast talk show, where the speaker knew the topic in advance and was a practiced orator. The difference between 0.70 and the highest result from the participant data is roughly equal to the range of the participants' MATTRs for a given time period, and considerably larger than the difference between an individual participant's MATTRs in 2009 and 2010. This exploratory result seems to indicate that MATTR (with an appropriate window value) is promising in terms of discerning between intermediate and advanced (or native) proficiency.<sup>23</sup>

---

<sup>22</sup> I used an excerpt from a discussion of housing issues on the weekly CCTV program "Dialogue" 《对话》 held on December 13, 2009. The sample chosen was the first 650 words of the main guest's turns to speak, not including prefatory remarks and chat. (Retrieved from <http://space.tv.cctv.com/article/ART11260925239708652> and <http://space.tv.cctv.com/video/VIDE1260722337607885> on March 28, 2012.) I chose this because (1) the audio and the transcript were already available; (2) the speaker was a well-educated adult, similar to the study participants; (3) the subject matter (comments on life in China and Chinese society) was similar to the questions posed in parts of the participant interviews in this thesis (though as the main guest was the mayor of Harbin, his command of the subject matter could be considered much stronger than the participants examined in this thesis).

<sup>23</sup> I also attempted to find transcripts of Chinese child speech as another indication of less-varied speech to compare to the participants' learner speech. Of the resources I could find through the internet, three looked promising but turned out to be inadequate for the purposes of this study:

(1) The CHILDES project (<http://childes.psy.cmu.edu/>) only provided transcripts in romanized format, which the parsers used in this thesis are not capable of analyzing;

(2) SCoRE (<http://corpus.nie.edu.sg/wordlist/index.htm>) was not freely available; and,

(3) CanCorp (<http://www.arts.cuhk.edu.hk/~lal/index.php?id=9>) consisted of participants and data that were difficult to compare to the present study's. (The eight children whose speech is transcribed are all very young, under four years of

## 4.2.2 Lexical sophistication

For the study I examined three measures of lexical sophistication: **sophisticated TTR** (an indication of the level of sophistication of words used; see 2.2.4.1 and 2.3.1 for further description of this measure), **polysemy** (the variety of meanings employed for a given morpheme), and **mean word frequency** (a rough estimate of the propensity for less-common lexical use).

### 4.2.2.1 *Sophisticated TTR*

Based in part on the research group's ideas and in part on Jun Da's (2005, 2006) experimental Chinese Vocabulary Profiler 1.0, I calculated **sophisticated TTR** measures in two ways, comparing the unique words used in each interview to:

1. The 5,000-most-common tokens in the **Lancaster Corpus of Mandarin Chinese (LCMC)**<sup>24</sup>, for an indication of the frequencies of the words used by each participant (henceforth LCMC-5000); and,
2. The contents of the six-level **HSK word list**<sup>25</sup>, for an indication the pedagogical level of each participant's word choices.

---

age, and thus the dialogs were not conversational; there was not enough speech generated by the children in any one turn to be comparable to the conversational parameters of the current study.)

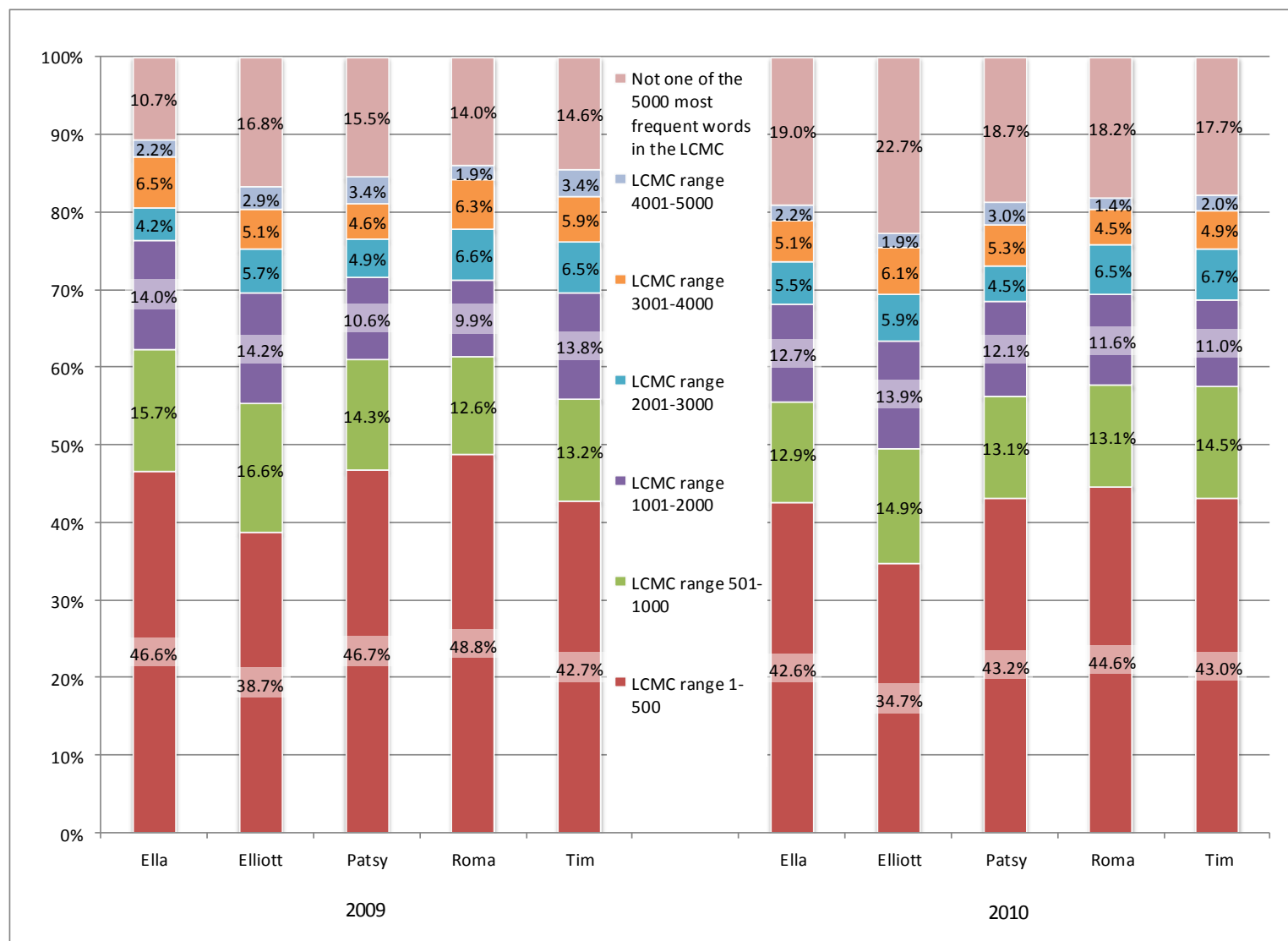
<sup>24</sup> The LCMC-5000 list, of the 5,000 most common tokens appearing in the balanced corpus of varied Mainland Chinese written texts comprising the LCMC, was downloaded from <http://corpus.leeds.ac.uk/frqc/lcmc.num> on August 3, 2011. This list was chosen because it was freely available, it was in a format conducive to scripted analysis, and its origins were explained by the compilers.

<sup>25</sup> The HSK word list, released by Hanban as a reference for students and educators in preparing for the revised 2009 HSK exam, was downloaded from <http://popupchinese.com/hsk/test> on August 6, 2011. The resources provided by popupchinese.com contain a total of 4,982 entries.

#### 4.2.2.1.1 *Calculations and observations*

For the word frequency calculations, I computed the percentage of unique words used by each participant that fell within different frequency ranges as referenced to the LCMC-5000 list. The results of these calculations for each participant's 2009 and 2010 interviews are shown in Figure 4-3. I then compared the unique words used by each participant with the HSK word list, the results of which are shown in Figure 4-4. Note that in both cases, as mentioned above, the percentages given are relative to the unique words, or vocabulary, used by the participants, independent of the frequency of any given word's use throughout an interview (frequency of word use is examined in section 4.2.1, above, in the context of lexical variety).

The general idea of such analysis is straightforward and intuitive, though interpretation of the results can be problematic; first I will present the results, then I will examine many of the issues that arise on closer inspection. Consider the implications of the vocabulary frequency data as shown in Figure 4-3.



**Figure 4-3: Percentage of words from interviews falling into different frequency ranges of Lancaster Corpus of Mandarin Chinese (LCMC) 5000-most-frequent token list (LCMC-5000) in 2009 and 2010**

We would expect that a higher percentage of low-frequency vocabulary would appear in the speech of advanced-proficiency speakers, and that conversely a greater proportion of lower-proficiency speaker word choices would fall in the high-frequency range. On the chart, the top-most categories (nearing 100% on the vertical axis) are low-frequency words, and the bottom-most categories (nearing 0%) represent high-frequency words. We note that:

- Relatively few of Ella's 2009 lexical choices were outside of the LCMC-5000, whereas in 2010 she showed the largest increase in that category. In 2010 the percentages of each of the different bands in the LCMC-500 all went down excepting the 2001-3000 range.
- Elliott showed higher-than-average increases in the 3001-4000 range and for words not on the LCMC-5000 list.
- Patsy's higher-than-average increases were in the 1001-2000 range and the 3001-4000 range.
- Roma's higher-than-average increases were in the 501-2000 range.
- Tim's higher-than-average increases were in the 1-1000 range.
- There is a consistent trend for all participants towards using more words outside the LCMC-5000 in 2010 compared to 2009; the average increase is 5.0%, with the majority of that change reflected in the 3.1% average decrease in use of words in the 1-500 range.

We could posit that this last trend is due to the fact that two pictures were shown in the 2010 interview, rather than just one in 2009, providing more opportunity for unusual words, or perhaps that the participants were on average more lexically expansive in the second interview due to familiarity with the interview process, the expected answer format, and each other.

Next let's examine the HSK results in Figure 4-4.

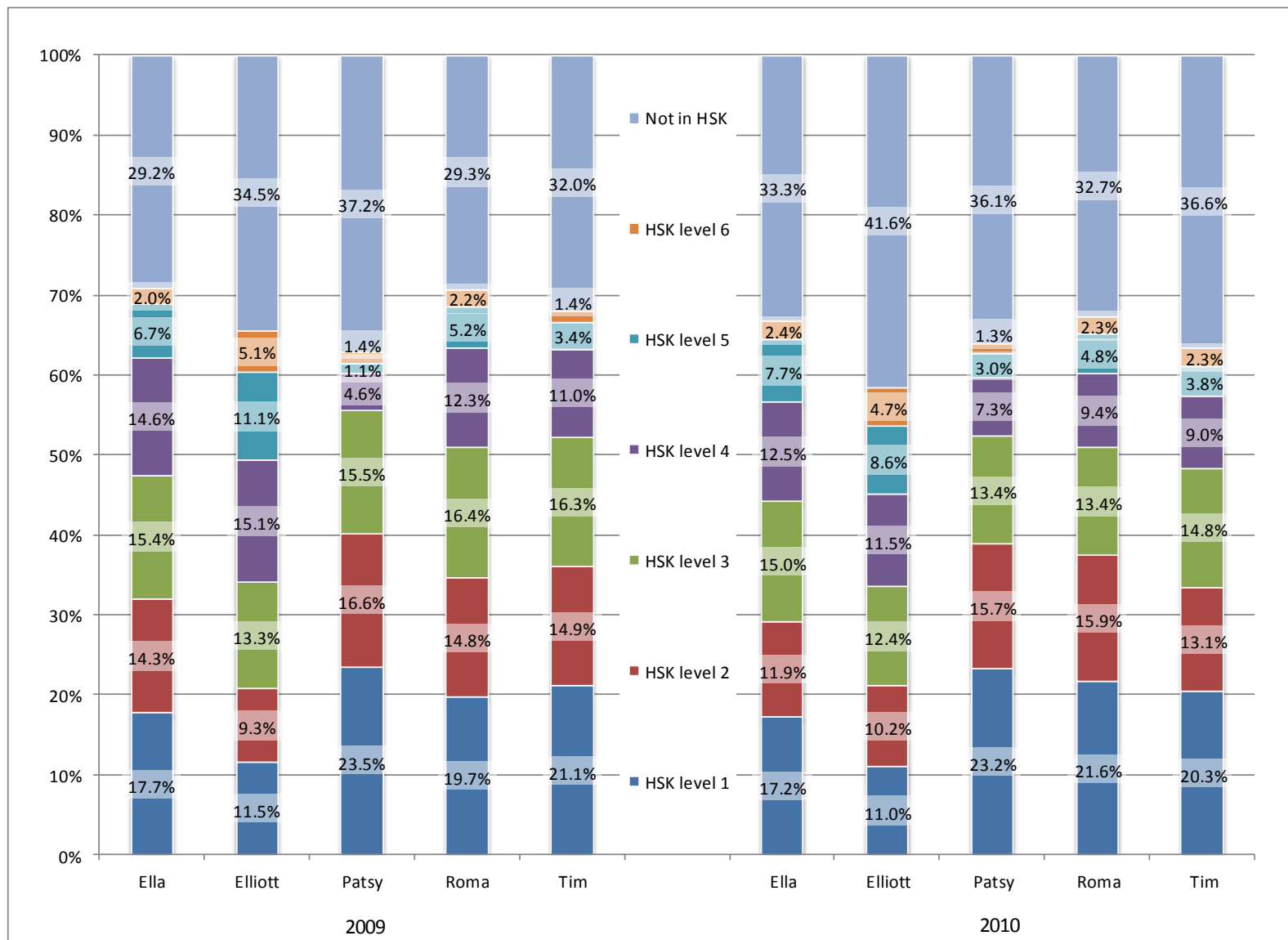


Figure 4-4: Percentage of words from interviews falling into different HSK vocabulary levels in 2009 and 2010

We can make the following observations:

- In 2009, Elliott's HSK 1-3 vocabulary made up 38.1% of his speech, and HSK 4-6 vocabulary accounted for another 31.3%. This relatively high percentage of HSK 4-6 vocabulary is consistent with his 2009 CEFR self-assessment in oral proficiency, the highest of the group at B2.
- Also in 2009, Patsy and Tim rated themselves at A2 for oral proficiency. Patsy's HSK 1-3 was indeed the highest portion in the group, and HSK 4-6 the lowest; Tim's HSK 1-3 percentage, however, was closer to Roma's (who assessed herself at B1) than Patsy's, and his HSK 4-6 percentage was much closer to Roma's and Ella's than Patsy's, indicating a qualitative difference in the types of language Patsy and Tim were using while both assessing themselves at A2. A closer look at the vocabulary they were using reveals that Tim was employing more sophisticated language:
  - HSK 6: Tim used 偏见 "prejudice", 摇滚 "rock and roll", 毒品 "drug", 非法 "illegal"; Patsy used 比方 "for example", 嗯 (*ng*, interjection / onomatopoeia "un / ng"; as a written form for a common sound, this is likely on the HSK list due to its low frequency in written texts, and the need to test learners in character recognition, as opposed to any sophistication accorded to its use in speech), 端 (*duān*, in this case a proper name; it appears on the HSK list as a noun, "end, tip," and thus this occurrence is not indicative of sophisticated language use, but of producing a name);
  - HSK 5: Tim used 餐厅 "restaurant", 进步 "advance, progress", 论文 "academic paper, thesis", 老板 "boss, owner", 移民 "immigrant", 理由 "reason, cause", 未来 "future", 期待 "to expect, to await", 报告 "report", 差别 "difference", 完美 "perfect", 哲学 "philosophy"; Patsy used 营养 "nutrition", 竹子 "bamboo", 拿 "pick up, take", 不好意思 (*bùhǎoyìsi*, expressing shyness, apology, or embarrassment);

- HSK 4: Patsy used 11 unique HSK level 4 words; she used these words a total of 16 times in her interview (therefore she used some words just once, and some more than once). Tim used 34 unique HSK level 4 words a total of 67 times in his interview (again, some words just once, some multiple times, such that there were 67 occurrences of words from this set of 34 unique words in his interview transcript).
- In 2010 both Ella and Tim self-assessed at higher levels than in 2009. Their HSK 1-3 vocabulary percentage shrunk somewhat, by 3-4%, and though the HSK 4-6 vocabulary percentage did not change much, the amount of vocabulary used that was not on the HSK list increased; as we shall see below, it is quite difficult to interpret these results using the current method.
- Patsy's 2010 numbers were higher for HSK 4-6 than 2009, which perhaps points to a real gain in sophistication despite her feeling that her oral proficiency remained at the A2 level. This could indicate that despite expressing herself better in 2010, she may still have felt that the A2 criteria represented her functional capabilities due to the functional focus of the 'can do' criteria on the CEFR; if so, it points to the potential viability of considering lexical sophistication as somewhat separate from pragmatic or functional proficiency.<sup>26</sup>

The LCMC-5000 results are not very clearly indicative of CEFR level, but the HSK results do appear a little more comprehensible. However, digging in a little deeper reveals there are serious faults with a facile comparison of the interview 'words' as determined by ICTCLAS and the words on the LCMC-5000 and HSK list.

First, the LCMC-5000 list is simply a frequency-ordered account of the tokens found in the LCMC; the segmentation process was not perfect. Indeed, the 5000-word reference data includes many tokens that are

---

<sup>26</sup> For examples of functional criteria at the A2 level, see Table A-2 and Table A-3 in Appendix A.



not strictly comparable to actual speech data (e.g. there are four instances of punctuation in the 25-most-frequent tokens, and elsewhere in the list can be found specific orthographical choices for representing numbers and dates that may not align with equivalent phrases transcribed in different ways; it also contained some multiple-word phrases such as 科学技术 ‘science and technology’).

Second, the HSK word list was created for pedagogical purposes, so its contents differ greatly from both some of the tokens generated by ICTCLAS’s segmenting principles and those of the LCMC-5000. It is perhaps easiest to see this difference by looking at the actual tokens participants used in their interviews (as parsed by ICTCLAS) that were not on the LCMC-5000 or HSK list, shown in Table 4-1, and discussing the difficulties in interpreting these examples as highly sophisticated lexical choices.

**Table 4-1: ICTCLAS-segmented ‘outlier’ tokens from interviews (not on the LCMC-5000 or HSK list)**

OK	乡下	加拿大	图像	小孩子	教书	河边	男朋友	翻过	这边
儿	乡下人	北美	图片	小帽	教课	油饼	画家	老外	这部
一两	乱七八糟	北美洲	土耳其	小时候	整合	法文	画画	老朋友	逻辑
一两年	二二三	北部	圣诞节	小朋友	文书	法语	疏远	职员	那段
一九三零年	二十二十	十五	坏事	小路	文法	泥土	瘟疫	联邦	邱晓龙
一九九九	二十年	十年	坏人	尼姑	文言文	注意力	登山	聘用	部份
一九二二九	二月	十月	坏处	山上	料理	泰	白人	聘请	部落
一九零	五九年	升级	埃及	巴士	新书	泰国	百分之百	肉体	部首
一对一	五五	华语	墨西哥	师范大学	新加坡	泰文	百货大楼	肯尼亚	都用
一年	五分	单字	墨西哥人	庞	新区	洲	皈依	育	重金属
一点儿	五年	单词	外婆	庞克	旅行	派对	省事	自传	金边
一点点	五月	卖淫	外来	庞克人	无所谓	济	看不到	自然环境	钓鱼
一百五十	五百	南部	外省人	度假	日文	海滩	看中	荷兰	钟头
七十	交朋友	卡尔加里	多久	建立不同	时时	渥太华	看电视	蒙古	钢琴
三三	交通系统	印尼	多伦多	建筑物	明尼苏达	温哥华	看病	蓝色	长春
三十三	亲密	厦门	大山	开始工作	明尼苏达州	滥用	短期	补习班	闹钟
三四	亲近	发现新	大楼	开车	明年	演	短篇	西双版纳	阿拉
三月	人际	发音	大理	弯曲	明白人	潜能	砵	西方人	阿拉伯语
三百二十六	价钱	叙	夸张	影响力	星期六	点菜	破旧	西班牙	陆线
上映	伊斯兰	可不可以	奖学金	很难说	星期天	烤	硕士生	观察角度	雇用
上次	低调	合肥	女朋友	徒	星期日	烦	礼拜	警察局	露天
下课	保罗	吉他	好久	快要	替代	烫伤	社区	讯息	静坐
不在乎	信奉	同屋	好人	念书	有意思	热衷	福建	讲台	静安区
不太	做饭	听力	好奇心	性感	有时候	煮饭	私	讲述	非常少
不对	健步	听听	好玩	恩	有用	爱人	私立	诉	非洲
不是	傣	吸毒	好玩儿	悠闲	李白	父母亲	移居	试试	面试
与众不同	傣族	吸烟	好看	惨	杜甫	爸	穷人	诗词	韩国
专用	充分发挥	呃	好笑	懦弱	柬埔寨	王之涣	章鱼	课堂	韩文
东亚	全天	咖啡店	孙女	成中文	栋	王维之	简讯	贫穷	韩语
东海	全球性	哈哈	学习者	成就感	格	玩儿	簪	足球	音乐家
两两	公众	哎呀	孩	我教	棘手	珈	精采	足迹	预备
两年	公费	哥伦比亚	安全感	手臂	正统	珠海	糟	路程	飞机场
中华	六十年代	唐诗	安徽	打字	每周	球技	索马里	路边	餐巾
中型	内蒙古	商界	定居	打鼓	比如说	球队	繁多	车子	饼
中外	写字	喔	实时	技	民社	瑜	纪录片	转向	高雄
中级	写法	喝茶	室友	拜拜	汇	甜	纸张	轻易	高高
中英英	农夫	囉	家族	探险	汉人	生字	纹	辑	麦当劳
主角	凝	四季	宿	摩	汉堡	生词	网上	辞	麻将
乐团	出事	团队	宿心	政治学	沙特阿拉伯	电线	网路	迎合	黑人
习	分居	国语	寺	故事情节	没什么	电视剧	网页	这么样	

#### 4.2.2.1.2 Issues

Note that there are many different types of ‘tokens’ in Table 4-1 that do not strictly match up with entries on the HSK list or LCMC-5000, but still appear to be commonly used; there are compound words, proper names, numbers, dates, and numerous multiple-word tokens that appear to have been erroneously judged by the segmenter to be a single ‘word.’ Thus it is quite easy to see that there are major issues in comparing the segmenter output to these two word lists and expecting clear results. I attempt below to explain many of these issues.

##### 4.2.2.1.2.1 Compound words or multiple-word tokens

Examples:

- 政治学 ‘political science, politics’
- 国语 ‘national language, Chinese’
- 网上 ‘online’

Many of the outlier tokens in Table 4-1 are compounds of basic morphemes that appear on the HSK list, which does not exhaustively include compounds resulting from basic morphemes as the possibilities are so numerous. Thus, words like 政治学 ‘political science’ and 国语 ‘national language, Chinese’ do not appear on the HSK list presumably because the authors felt they were adequately accounted for by the presence of 政治 ‘politics’ + 学 ‘the study of’, and 国 ‘country’ + 语 ‘language.’

Interestingly, these latter three single-syllable morphemes do not appear on the HSK list either; again, presumably the authors felt that their importance to learners and teachers would be evident in the inclusion on the list of words they appear in as bound forms, such as 科学 ‘science’, 数学 ‘math’, or 化学 ‘chemistry’; 国家 ‘country’ or 国际 ‘international’; and 汉语 ‘Chinese’ or 母语 ‘mother tongue, native language’. Given each of these two-syllable words fall under different HSK levels, it is not a trivial matter to determine what level of

sophistication the authors of the HSK consider exemplified by use of the constituent forms in compounds that do not appear on the list.

#### 4.2.2.1.2.2 *Proper names*

Examples:

- 明尼苏达州 ‘Minnesota’
- 杜甫 ‘Du Fu’
- 傣族 ‘Dai (peoples)’

Proper nouns would not be listed exhaustively on proficiency word lists, as they are situational to learning context and individual learner experience, and potentially infinite in number. While they may be on raw frequency lists for corpora, they are of course dependent on the corpora content; for the purposes of examining lexical sophistication they might be considered ‘noise.’

Again, this does not appear to be a hard-and-fast rule for HSK list designers, as some proper nouns do appear, e.g. 北京 ‘Beijing’ (which is more predictable in the Chinese-language-learning context than Minnesota or Phnom Penh , two places mentioned by participants in their interviews and appearing in Table 4-1).

#### 4.2.2.1.2.3 *Number constructions*

Examples:

- 六十年代 ‘the Sixties’
- 五五 ‘five five’ (repetition of a morpheme)
- 一两 ‘one or two’

The combinatorial possibilities of numbers are too varied to merit including them one-by-one on pedagogical word lists, thus the combinations as segmented by ICTCLAS do not represent sophisticated lexical use. For instance, take 六十年代 ‘the Sixties,’ where 六 ‘six’ and 十 ‘ten’ are common combinatorial morphemes, and 年代 ‘decade’ is an HSK 5 word. The use of this combination of morphemes is evidence of grammatical sophistication, but not strictly speaking lexical sophistication; ICTCLAS presumably considers it a ‘chunk’ with a single inherent meaning (similar to the English translation), whereas on a proficiency list these components would be taught separately and not repeated in their combinatorial possibilities.

Further, there are transcription issues involved as well. Numbers in Chinese can be digitally encoded in multiple ways, e.g. in half-width ASCII Arabic numeral form (e.g. 1), in Chinese numeral form (e.g. 一), and in full-width Arabic numeral form (e.g. 一 ). The LCMC-5000 lists some numbers in this latter form, whereas the transcription choice for the interview data used in this thesis was the Chinese numeral form; the scripts I wrote worked purely on comparing the digitally encoded forms, and were not ‘intelligent’ enough to determine that meaning-wise these characters were equivalent.

#### 4.2.2.1.2.4 Syntactic or orthographic variations of words that might otherwise appear on a proficiency list

Examples:

- 一点点 ‘just a little bit’
- 高高 ‘high, tall’
- 部份 ‘part’

These tokens are duplications and variations on words that might be found in pedagogical word lists, or they are measure-like compounds that would not be expected to be listed in all their possible forms on pedagogical word lists.

一 *yī* and 点 *diǎn* are both on the HSK list, but not 一点 ‘a little bit,’ and certainly not the duplicative 一点点 ‘just a little bit,’ which was determined to be a single word by ICTCLAS. Interestingly, the 点 that appears on the HSK list is for the meaning ‘o’clock,’ and not ‘little bit,’ which raises the thorny problem of how to distinguish whether a token like 一点点 actually does correspond to basic meanings in the reference list; the constituent characters of the token in question might match up with characters on the list, but the meanings might not, thus it would be necessary to examine these on a case-by-case basis.

高高 duplicated ‘high, tall’ is not a lexical compound but rather a syntactic compound, the duplication of a single-syllable adjective for rhetorical effect. This might be considered syntactic complexity rather than lexical sophistication; its constituent morphemes are found on the HSK.

部份 ‘part’ is an orthographical variant of 部分 ‘part;’ the latter form is the HSK version, while the former does not strictly appear on the HSK because of the variation in the second character. This occurred possibly as the result of a transcription choice or input method error, and highlights the difference between human comparison of tokens with reference lists (which would likely decide that the two terms were equivalent) and computer comparison (which merely sees the second character in each as different and thus does not match the two together).

#### 4.2.2.1.2.5 Oral language

Examples:

- 玩儿 *wánr* ‘to play’
- 一点儿 *yīdiǎnr* ‘a little bit’
- 烦 ‘annoying’
- 哎呀 (onomatopoeic exclamation)
- 哈哈 (the sound of laughter)

Some transcribed words were tokenized by ICTCLAS but do not appear on the HSK list or LCMC-5000 due to their intrinsically oral characteristics; that is, they represent words or pronunciations that occur more frequently in informal speech than in writing. For instance, 一点儿 ‘a little bit,’ which adds an 儿 *ér* to 一点 ‘a little bit’ to denote that the speaker used a rhotacized ending, pronouncing the term *yīdiǎnr*, is essentially the same word with a different quality, but again computerized comparison would not match it with the equivalent word on a reference list automatically (in this case, on the LCMC-5000 the written form 一点 ‘a little bit’ is #377). Some expressions are essentially oral and need to be transcribed in some way, but may not be transcribed in a way common in mixed texts (LCMC-5000) or pedagogic texts (HSK). 哎呀 (onomatopoeic exclamation) and 哈哈 (the sound of laughter) are two such examples. Finally, some lexical choices may be exclusive to speech or to literature that attempts to transcribe speech; for instance, 烦 ‘annoying’ is common in some topolects of Chinese, but the various entries in the HSK word list that include 烦 are not equivalent (though it could arguably be considered a short form of the HSK word 烦恼 ‘worried’).

It is again worth noting that some examples of this type of word are found on the HSK list, e.g. 哪儿 *nár* ‘where...?’, but once more the possible combinations are so many that presumably the authors did not attempt list more of them.

#### 4.2.2.1.2.6 Segmenting errors

Examples:

- 习 *xí* ‘to practice’
- 建立不同 ‘establish’ + ‘different’
- 可不可以 ‘can [I]...?’
- 看不到 ‘cannot see’
- 开始工作 ‘start working’
- 很难说 ‘difficult to say’

There are many cases where the segmenter did not act in a way that is coherent with how a human would interpret the text. In some cases words were broken into single morphemes due to punctuation or ambiguous context (e.g. 习 *xí*, a part of 学习 ‘learn,’ isolated due to punctuation; note that as mentioned in 3.3.1 the punctuation was removed after the parsing step). In other instances the segmenter tokenized groups of words into ‘chunks’ which might be more appropriately broken down into grammatically separate constituent parts: e.g. 建立不同 ‘establish’ + ‘different’ should really be two words; 看不到 ‘cannot see’ is a verb followed by a resultative complement.

#### 4.2.2.1.3 Discussion

It appears that it would be a formidable task to develop scripts that would take into account these issues and effectively map the segmented data onto the expectations of the other reference data sets. It is still possible to observe lexical use which does imply sophistication beyond the specific examples in the LCMC-5000 or HSK list, e.g. from Table 4-1: 与众不同 ‘extraordinary’, 乱七八糟 ‘a real mess’, 滥用 ‘to misuse’, 懦弱 ‘cowardly’, 亲密 ‘intimate’, 预备 ‘preparation’, and 课堂 ‘classroom.’ Further, there are ‘chunks’ which might be considered lexicalized combinations indicative of comfortable, fluent, natural speech but not strictly single words, such as 比如说 ‘for instance’ and 无所谓 *wúsuǒwèi* (expressing indifference); it remains ambiguous



whether or not these should be considered sophisticated lexical use or merely combinations of more basic components (note that HSK level 1 contains the phrases 没关系 ‘it doesn’t matter’ and 不客气 ‘you’re welcome’).

In any case, these sophisticated outliers represent a distinct minority among the tokens listed in table 4-1. Most of the outlier tokens are only explicable as low-frequency occurrences due to the nature of (1) the segmenting tools, and (2) the strict use of reference lists, and thus do not necessarily represent highly sophisticated lexical use. Segmenting programs use statistical analyses of reference corpora, models of speech functions and relationships, and probabilistic heuristics to make a best guess on how to consistently split up novel sentences into coherent and grammatical tokens (‘words’), but they are not always perfect. Their decisions may be grammatically questionable (as in the last category of examples given above), or simply different from the expected divisions (e.g. bundling simpler morphemes into ‘chunks’ if commonly seen together, or keeping rhetorical doubling of verbs and adjectives together as a single ‘word’, etc.).

The idea of what divisions are indeed ‘expected’ segues nicely into the second explanation mentioned for many of the outlier words, the reference lists: these lists are built for specific purposes and different genres of text may find more or less overlap in terms of vocabulary frequency (oral vs. written) and word forms (orthographic representations, regional differences, or the vast combinatorial variety of numbers, dates, and names, for instance).

As these outliers do not unambiguously represent ‘high-level language use,’ it is perhaps more interesting to observe the use of words that lie within defined HSK ranges. Keeping in mind the caveat already mentioned above concerning the rigid nature of reference lists, the HSK vocabulary results that do fall within levels 1-6 do appear more promising in correlating with proficiency, and in some cases may help distinguish learners who are otherwise assessed at the same level on a scale with broad bands such as the CEFR. The self-assessments were holistic scores, and presumably took into account ideas of fluency, comfort, speed, and situational/functional interaction in addition to lexical sophistication, so using more precise techniques to charac-

terize the range of lexical choices may help coax out relationships between broad categories and concrete exemplars, and help characterize language use that might otherwise all be categorized in a single proficiency bloc. That said, if the outlier tokens in Table 4-1 were attributed to core vocabulary in HSK levels 1-6, that would change the relative ratios of each participant's HSK 1-6 use (as summarized in Table 4-1), so the results in this thesis must be assumed to be rough estimates without further manual validation. Though outside the scope of the work done in this thesis, some possible improvements to this method of lexical sophistication analysis are discussed in 5.3.1.2.

#### **4.2.2.2 Polysemous morphemes**

The research group together with Dr. Duanduan Li also wondered if exhibiting good knowledge of the multiple meanings of **polysemous morphemes** might imply a higher proficiency level, and to that end we brainstormed to identify some morphemes that might be helpful to quickly and clearly reveal a range of meanings and usages, some of which might be associated with a higher level of proficiency than others. It was decided that morphemes of movement would be a good start, thus I examined use of the polysemous movement morphemes 到 *dào* and 来 *lái*, which have basic meanings related to physical motion, but can be used in many other ways as well, e.g. in directional complements, abstract complements, and bound forms (where the movement morpheme has been essentially lexicalized as an integral part of a multi-syllabic word). I used the search feature of a word-processing application to look at all instances of the morphemes and coded each instance as a specific grammatical usage, then compiled all the discovered categories into Table 4-2 and Table 4-3, which indicate the presence of accurate use of a morpheme in a given way for each participant.

**Table 4-2: Participant use of the morpheme 到 *dào***

	Verb <i>to arrive</i>	Coverb (location)	Coverb (time)	Resultative complement
Ella	✓	✓	✓	✓
Elliott		✓	✓	✓
Patsy	✓	✓	✓	✓
Roma		✓		✓
Tim	✓	✓	✓	✓
Examples of usage	到西方的 荷兰以后 <i>dào xīfāng de Hólán yǐhòu</i> 'after ar- riving in Holland in the West'	搬到那里 去 <i>bān dào nàlǐ qù</i> 'move (to) there'	十岁到二 十岁 <i>shí suì dào èrshí suì</i> 'ten to twelve years old'	得不到 <i>dé bu dào</i> 'cannot get';  看到 <i>kàn dào</i> 'to see'

**Table 4-3: Participant use of the morpheme 来 *lái***

	Verb <i>to come</i>	Bound forms	Directional complement	Abstract complement	Set phrases	Coverb (purposeful)	<i>Since, continuing</i>
Ella	✓	✓	✓	✓		✓	
Elliott	✓	✓	✓	✓	✓	✓	✓
Patsy	✓	✓	✓	✓			
Roma	✓	✓					
Tim	✓	✓		✓	✓		
Examples of usage	不知道他 们什么时 候来 <i>bù zhīdào tāmen shénme shíhòu lái</i> 'don't know when they're coming'	从来 <i>cónglái</i> 'all along'  来自 <i>láizì</i> 'come from'	拿来 <i>nálái</i> 'bring'  回来 <i>Huílái</i> 'come back'	想不起来 <i>xiǎng bu qǐlái</i> 'can't think of [it], can't remember'	对我来说 <i>duì wǒ lái shuō</i> 'in my opinion'  越来越多 <i>yuè lái yuè duō</i> 'more and more'	中文是给我 其他的东 西来看 世界 <i>Zhōngwén shì gěi wǒ qí tā de dōngxi lái kàn shìjiè</i> 'Chinese gave me more ways to view the world'	这两年来没 有很多机 会 <i>zhè liǎng nián lái méiyǒu hěn duō jīhuì</i> 'haven't had many oppor- tunities for the past two years'

We can see that of the two, 来 *lái* is the more discerning for this data set, as a greater variety of uses was found (i.e. coded). Elliott showed the most variety of uses of 来 *lái*, followed by Ella, Patsy and Tim, and Roma. Interestingly, this does not exactly correspond to the participants' relative reported oral proficiencies (Patsy reported the lowest ranking at A2; Elliott and Ella reported similar rankings converging at the B2 level). In the case of 到 *dào*, little difference was observed in the variety of uses and gradient of sophistication in uses amongst the participants; it is possible that this particular morpheme is less discerning as it has less meanings, and therefore an easier-to-master range of polysemy.<sup>27</sup>

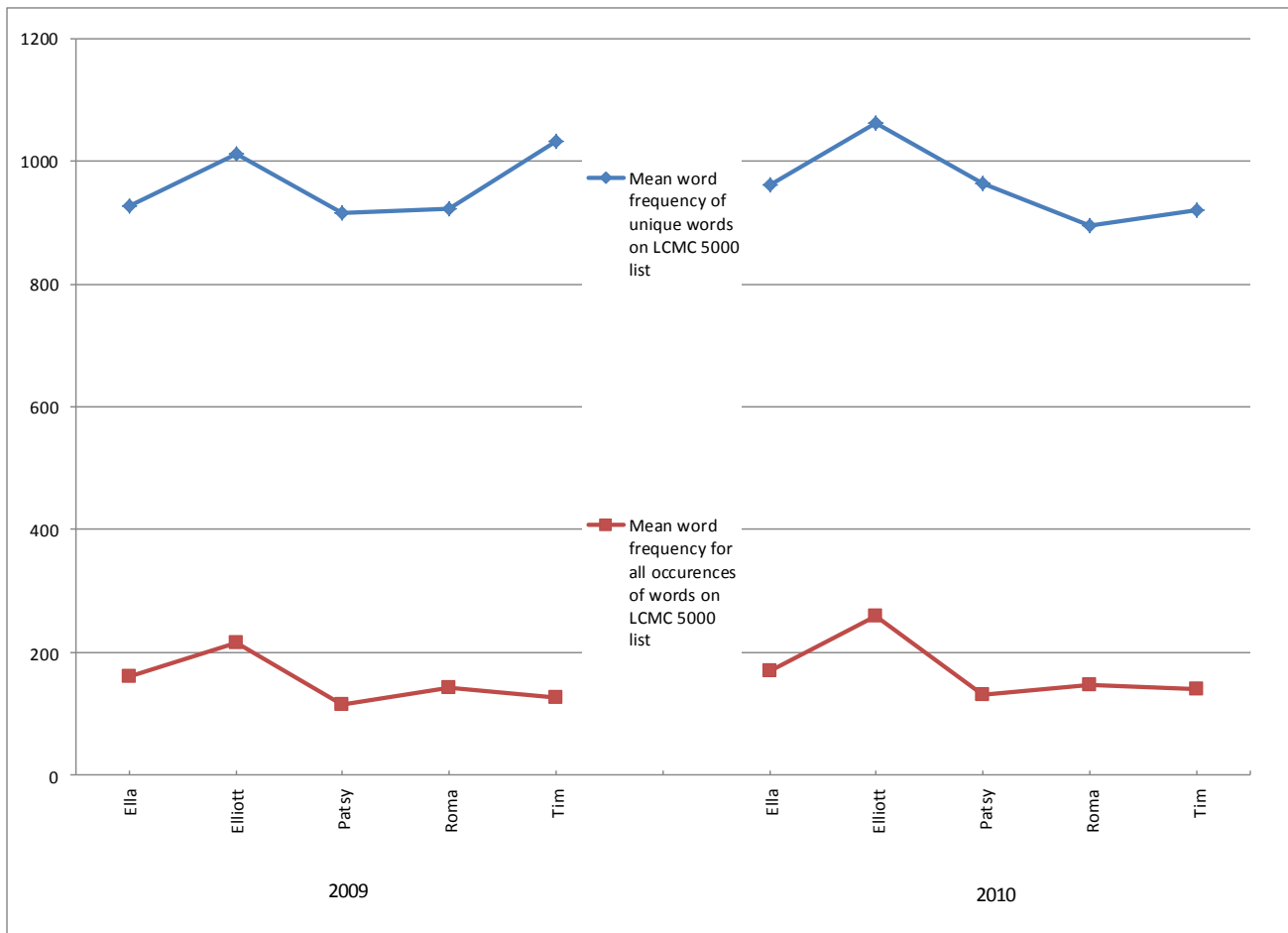
Though perhaps not enough data is available here, and of not a reliable enough calibre to discern the CEFR levels A2 to B2, this technique still shows one possible way to characterize lexical sophistication, particularly if it could be expanded to include more polysemous words, with more consistent coding. Also note that only grammatical appearances of these morphemes were included in the analysis, so the aspect of accuracy is not included, nor is the full range of possible meanings for each morpheme included.

#### 4.2.2.3 *Mean word frequency*

Finally, **mean word frequency** was calculated by averaging all of the unique words for which there were defined LCMC word frequencies; the results are shown in Figure 4-5.

---

<sup>27</sup> In Zhonghua (2004), 来 *lái* is listed with seven separate meanings (not including its use as a family name), whereas the entry for 到 *dào* lists just three meanings; this would support the observation of a wider variety in participant use of 来 *lái*.



**Figure 4-5: Mean word frequency of participant speech with reference to LCMC-5000**

The upper line in Figure 4-5 is a different way of representing the results shown in Figure 4-3 above, essentially that in some cases of the words that can reliably be tagged with a frequency ranking, some participants used on average more words of lower frequency (higher score in Figure 4-5). The differences appear to be relatively small, but that is in part due to the discernment of the reference data in this study: the LCMC list consisted of 5000 words, so any less-frequent words were discounted from the calculations presented in Figure 4-5 (these correspond to all words listed captured by the ‘outside LCMC 5000 range’ in Figure 4-3).

One interesting thing to note in Figure 4-5 is the difference in Tim’s 2009 data between mean word frequency taking into account all words spoken vs. just unique words (much higher than the ‘all words’ meas-

ure). That is, of the unique words (a word only counted once no matter how many times reused in the interview data), on average their frequency rankings were relatively high (they were ‘rarer’ words than those used by the other participants); however, considering the much lower ‘all words’ score, Tim must have used more-frequent words a great deal more often than less-frequent words, bringing the average word frequency score down.

### 4.2.3 Syntactic complexity

Though some researchers have found syntactic complexity measures useful in establishing the statistical significance of differences observed in proficiency (Jin, 2007), it is a difficult measure to operationalize as it typically requires careful manual coding. The procedure for parsing learner language into production units is onerous, requires a careful and informed distinction in grammatical categories, and may require inter-rater reliability checks, which together means significant time and expertise is required for such ratings.

Given this study’s focus on what computer tools can do to aid the linguistic analysis of Chinese as an Additional Language, I explored the capabilities of two freely available computational linguistic parsing libraries, ICTCLAS (introduced above, in section 3.3.1) and the Stanford Parser (henceforth SP).<sup>28</sup> Both are able to do some automated grammatical analysis, tagging tokens from a source text with varying degrees (and accuracies) of grammatical metadata. For the sake of initial investigation, and the inherent complication of grammatical parsing, I tried a simple few sentences from one interview (Elliott’s 2010); the detailed output from ICTCLASS and SP is given in Appendix B, and some observations are summarized here.

The example text used, as transcribed:

---

<sup>28</sup> The Stanford Parser (<http://nlp.stanford.edu/software/lex-parser.shtml>) is developed by the Stanford University Natural Language Processing Group; its grammatical parsing is based on the Penn Chinese Treebank and Stanford’s own typed dependencies analysis. Data for this thesis was obtained from the online demo (<http://nlp.stanford.edu:8080/parser/>) on April 8, 2012;

我...我目前正在学习的是...Berkeley 大学所提供的...um...一堂网上...中级...韩国话课。所以，我快要学完，但是学完这个之后，不是可以讲出来的，能...会讲出来的。而是看得懂就好。

“I... [the one] I’m currently studying is an intermediate Korean course offered by [UCLA] Berkeley. So, I’m almost finished, but after finishing it it’s not that I can speak it [*i.e. produce it out loud*], will be able to speak it [*correcting the auxiliary verb*]. Rather, it’ll be fine if I can just understand it when I read it.”

Note that this text includes:

- A single coherent topic, broken into multiple sentences which are related to one another
- Subordination and coordination
- Mistakes, repetition, and self-correction (e.g. 不是可以讲出来的，能...会讲出来的 “it’s not that I can speak it, will be able to speak it”)
- English words, as well as English letters representing non-speech sounds
- Punctuation, including two different forms of the ellipsis mark of three dots (indicating pauses in speech, not ellipsis)

ICTCLAS and SP are both capable of part-of-speech analysis (POS-tagging). The output from both libraries for the example sentence is shown in Table 4-4.<sup>29</sup>

---

<sup>29</sup> For a complete classification of the codes used in each system, please see [http://ictclas.org/ictclas\\_files.html](http://ictclas.org/ictclas_files.html) (for ICTCLAS) and <http://www.cis.upenn.edu/~chinese/> (for the Stanford Parser).

**Table 4-4: POS-tagging output from two grammatical parsers**

Parser	Example POS-tagging output
ICTCLAS	我/rr ./wj ./wj ./wj 我/rr 目前/t 在/p 学习/v 的/ude1 是/vshi ./wj ./wj ./wj Berkeley/x 大学/n 所/usuo 提供/v 的/ude1 ./wj ./wj ./wj um/x ./wj ./wj ./wj 一/m 堂/q 网上/s ./wj ./wj ./wj 中级/b ./wj ./wj ./wj 韩国/nsf 话/n 课/n 。/wj 所以/c ， /wd 我/rr 快/d 要/v 学/v 完/vi ， /wd 但是/c 学/v 完/vi 这个/rz 之后/f ， /wd 不/d 是/vshi 可以/v 讲/v 出来/vf 的/ude1 ， /wd 能/v .../ws 会/v 讲/v 出来/vf 的/ude1 。/wj 而是/c 看/v 得/ude3 懂/v 就/d 好/a 。/wj
SP	我/PN ./AD ./VV ./NN 我/PN 目前/NT 在/AD 学习/VV 的/DEC 是/VC ./NR ./JJ ./NN Berkeley/NR 大学/NN 所/MSP 提供/VV 的/DEC ./NN ./AD ./VV um/NN ./AD ./VV ./NN 一/CD 堂/NN 网/NN 上/LC ./AD ./VV ./NN 中级/JJ ./NN ./AD ./VV 韩国/NR 话/NN 课/VA 。/PU 所以/AD ， /PU 我/PN 快/VA 要/VV 学完/NN ， /PU 但是/AD 学完/NN 这个/NN 之后/LC ， /PU 不/AD 是/VC 可以/VV 讲/VV 出来/VV 的/DEC ， /PU 能/VV .../PU 会/VV 讲/VV 出来/VV 的/DEC 。/PU 而/AD 是/VC 看/VV 得/DER 懂/VV 就/AD 好/VA 。/PU

Clearly, this is difficult for a human to interpret at a glance, and the value in POS-tagging if used solely by a human coder is dubious, as the amount of ‘noise’ introduced by tagging each token, including punctuation, may override the convenience of automated classification of content and function words such as verb, adverb, and noun types, conjunctions, grammatical particles, and so forth. Nevertheless, computer scripts could be developed to harvest statistical data about parts of speech used. Here are just a few observations and ideas:

- **Ratio of verbs** (VA, VC, VV) to tokens in the speech: assuming regular, grammatical speech, this gives an indication of the complexity (number of subordinate terms or predicates) of each sentence. For instance, in the last phrase of the example sentence (SP output), we can see the predicate adjectival verb (labelled VA) occurs along with two other verbs and the copula, itself a verb as classified by SP, to form a complex sentence:
  - 而/AD 是/VC[verb-copula] 看/VV [verb-other]得/DER 懂/VV[verb-other] 就/AD 好/VA[verb-predicative adjective]



- **Verb chaining** can indicate complex auxiliary relationships between verbs. For instance, in the example sentence, we see a chain of verbs indicating auxiliary and resultative relationships to the main verb (also preceded by the copula):
  - 不/AD 是/VC 可以/VV 讲/VV 出来/VV 的/DEC
- Interesting cases of **particle use**, such as this example of the use of 所 to highlight the relationship of the actor to the action; a POS-tagger can pull out just these cases of particle use, ignoring cases where the particle may otherwise be performing a different grammatical function (e.g. 研究所 ‘research institute’):
  - Berkeley/NR 大学/NN 所/MSP 提供/VV 的/DEC
- We can imagine many other types of analysis that might be aided by POS-tagging:
  - **Measure word classification** (number of unique measure words used) as an indication of lexical variety (number of types) or lexical sophistication (for nouns or noun classes that admit multiple measure words depending on the situation or register, e.g. 一个人 ‘a person’, 一位先生 ‘a (gentle)man’, 一群学生 ‘a group of students’, 一伙土匪 ‘a company of bandits’).
  - **Variety of uses of number or amount** as an indication of lexical variety, or complex number combinations (十多个人 ‘more than ten people’, 十来个学生 ‘around ten students’) as an indication of syntactic complexity or syntactic variety.
  - **Adverb ratios** (e.g. adverb-to-verb ratios) may give an indication of syntactic complexity; this could be examined for other subordination or coordination such as **conjunctions** and **prepositions**.

SP’s POS tagger presents even more complicated capabilities, including recognizing **ba-constructions** and **bei-constructions** (把 and 被) only when used for these grammatical functions, aspectual particles (distinguishing these morphemes from their use in non-aspectual grammatical functions as well), and different uses of the *de* particle (complementizer, genitive, manner, aspect, sentence-final particle, etc.).

Note, however, that the SP POS-tagging output, which is richer than the ICTCLAS output, is not without errors. Here are a few examples:

- 学完 ‘study’ + ‘finish’ is marked as a noun (should be a verb + verb-resultative).
- 课 ‘class, course’ is marked as a predicative adjective (an adjective acting as the verb of the sentence) (should be a noun).
- Conjunctions (所以 ‘so’, 但是 ‘but’) are marked as adverbs.
- Punctuation and English are marked in bizarre ways; for instance, the string “um...” is marked as noun + adverb + verb + noun, clearly the result of a poor decision by the parser (should be perhaps marked as an interjection, an onomatopoeic particle, or even a foreign word)

An even more powerful function, but arguably correspondingly more difficult for human scanning and interpretation, is grammatical tree parsing, a function provided only by SP among the two libraries tested. Going beyond recognizing POS, this functionality parses text into coherent grammatical relationships, determining subject-predicate relationships, clause subordination, and so forth. Here is an example of the output of SP’s tree parser:

```
(ROOT
  (IP
    (NP
      (CP
        (IP
          (NP (PN 我))
          (VP
            (NP (NT 目前))
            (ADVP (AD 在))
            (VP (VV 学习))))
          (DEC 的)))
      (VP (VC 是)
        (NP
          (NP (NR .))
          (ADJP (JJ .))
          (NP (NN .))))))
```

This indicates that the hierarchical tree-branching relationships of the individual tokens to their functions in the text. For example, the string of text 我目前正在学习的 can be expressed as a {noun phrase}, further subdivided into {noun phrase – verb phrase – nominalizer}, and even broken up into individual tokens as {pronoun – time-noun – adverb – verb – nominalizer}; these various representations are equivalent and represent a nested grammar hierarchy.

The text that follows this segment, Berkeley 大学所提供的一堂网上中级韩国话课, is part of the same sentence, but the punctuation flummoxes the parser (in this case, the ellipsis “...” following 是; see Appendix B for the full output). With some manual pruning to ensure the parser doesn’t trip over the transcription choices used for the data in this thesis, we can use the following text, which the parser can handle much better:

我目前正在学习的是 Berkeley 大学所提供的一堂网上中级韩国话课。所以，我快要学完，但是学完这个之后，不是可以讲出来的，能，会讲出来的。而是看得懂就好。

“I’m currently studying an intermediate Korean course offered by [UCLA] Berkeley. So, I’m almost finished, but after finishing it it’s not that I can speak it [*i.e. produce it out loud*], will, will be able to speak it [*correcting the auxiliary verb*]. Rather, it’ll be fine if I can just understand it when I read it.”

The parser output for this text is rich with syntactic complexity information, a veritable gold mine. Without listing all of the results here, we can straight away see some data that could be analyzed to determine complexity (see Appendix B for further details):

- Sentences are broken into subject and predicate
- Subordinate terms are clearly indicated (though not entirely correct)

- Various categories of phrase such as noun phrases, verb phrases, locative phrases, and so forth are clearly marked.

In fact, with manipulation of the library's configuration (not possible in the demo version I used), functional information such as classification of verb objects (direct, indirect), focusing, imperatives, and even topic is possible. Due to the in-depth nature of the programming and the copious output, requiring development of scripted algorithms for analysis, I did not go into such detail in this thesis; possibilities will be discussed in Chapter 5.

It is clear that, given further software development, the time-consuming process of coding texts for grammatical categories and relationships in order to characterize the syntactic complexity of learner language can be reduced drastically. This is a promising subfield, but without further research into how to use the existing capabilities of such software libraries to consistently distill useful measures of syntactic complexity for CAL, the process of examining a text for complexity remains a difficult one. In addition, the accuracy of the POS tagging for oral speech is poor. For these reasons, I decided that known measures of syntactic complexity were not feasible to calculate for the full data set under investigation, as the amount of effort required to develop and verify a suite of test measures was beyond the scope of this thesis.

#### **4.2.4 Syntactic variety**

Given that traditional measures of syntactic variety do not map easily onto Chinese syntactic analysis (see section 2.3.1), the research group sought other indications of this measure. We decided upon examining the usage of key Chinese grammatical morphemes (similar to our exploration of polysemous morphemes as an indication of lexical variety; see section 4.2.2) to determine whether their presence, variety, or frequency indicated proficiency:

1. **Aspectual morphemes:** 过 *guò* (indicates experiential aspect) and 了 *le* (indicates perfective or completed aspect).
2. **Passive morpheme:** 被 *bèi*, used to indicate the passive.
3. **把 *bǎ* morpheme:** 把 *bǎ*, used to reposition the direct object in a sentence in order to focus on the object and/or highlight the result of an action.

The categories observed for aspectual morpheme use are summarized in Table 4-5 and Table 4-6 below.

**Table 4-5: Participant use of the morpheme 过 *guò***

	<b>Aspect: experience</b>	<b>Bound forms</b>	<b>Verb <i>cross, pass</i></b>	<b>Resultative complement</b>
Ella	✓	✓	✓	✓
Elliott	✓	✓	✓	
Patsy	✓			
Roma	✓			
Tim	✓	✓		
Examples of usage	没有去过 <i>méiyǒu qùguò</i> 'have never gone'	难过 <i>nánguò</i> 'sad'	过河 <i>guò hé</i> 'to cross or pass (a river)'	转过来 <i>zhuǎn guòlái</i> 'to turn around, turn over'

**Table 4-6: Participant use of the morpheme 了 *le***

	Aspect: completion	Aspect: change	Exclamation ... 了!	Bound forms	Other patterns
Ella	✓	✓	✓	✓	✓
Elliott	✓	✓	✓	✓	✓
Patsy	✓	✓	✓		
Roma	✓	✓	✓	✓	
Tim	✓	✓	✓		✓
Examples of usage	学了一点 <i>xuéle yīdiǎn</i> 'studied a little'	都没有了 <i>dōu méiyǒu le</i> 'none of them are there anymore'	有的人用太多了 <i>yǒude rén yòng tài duō le</i> 'some people use [it] too much!'	为了外国人表示这是新的中国 <i>wèile wàiguórén biǎoshì zhè shì xīn de Zhōngguó</i> 'to show foreigners that this is a new China'	有了手机才会 <i>yǒu le shǒujī cái huì</i> 'only with a cell phone'

Similar to the analysis performed on the morphemes 到 *dào* and 来 *lái* in section 4.2.2, we can see that at higher levels of proficiency (Ella and Elliott at B2), a wider variety of uses of the aspectual morphemes 过 *guò* and 了 *le* emerges. With further refinement of such individual 'scales' (e.g. looking at particular morphemes) of syntactic variety, it might be possible to develop a useful suite of individual indicators that combine to provide more nuanced distinctions of proficiency.

The results for *ba*- and *bei*-constructions were less discerning than was hoped. Ella, Elliott, and Roma did use the *ba*-construction, but just 1.5 times on average per interview (an average over both interviews); Patsy and Tim did not use it at all. Ella used the *bei*-construction four times in total over both interviews, and Elliott and Patsy used it just once; Roma and Tim did not use it in their interviews. It is tempting to ascribe higher-frequency use of *ba*- or *bei*-constructions to higher oral proficiency, given that Ella and Elliott (who assessed their oral proficiency in the B2 range) were the only participants to use both constructions, but with so few instances of use it is difficult to feel confident about any such conclusions.

Although not attempted for this thesis, grammatical parser output could also be examined for CAL input data for evidence of syntactic variety, as mentioned above in section 4.2.3. Such automated parsing of speech

uses, when combined with computer scripts to summarize results or make the results easier to confirm by human coders, might mean a greater set of syntactic relationships can be examined at once to develop a kind of global syntactic variety index; the full exploration of possibilities is beyond the scope of this thesis.

### **4.3 Accuracy measures**

No accuracy measures were chosen for this study. The primary reason was the research group's decision at the outset of the shared research to focus on the 'can do' aspect of our language capabilities; the data in this thesis is available for analysis only due to the research group's agreement on what it was to be used for and thus I did not attempt to validate or develop any of the accuracy measures discussed in 2.3.2.

In terms of their potential for automation, another of this thesis's guiding principles (discussed in 3.4), it must be noted that accuracy measures are time-consuming to code, and require excellent prescriptive knowledge of the language in question (including the context in which it is produced, for judgments of correctness); in addition, potential ambiguity and subjectivity arise when coding for accuracy. The accuracy measures discussed in section 2.3.2 do not lend themselves to easy operationalization via computer software. One way in which technology may still be useful in obtaining measures for accuracy is that given a list of trouble words or grammatical types to examine, the use of a grammatical parser along with a simple word processor can be invaluable in speeding up manual accuracy judgments by listing all instances of a given pattern in their context.

For further discussion of possible accuracy measures in the CAL context, see section 5.3.2.

### **4.4 Fluency measures**

Basic speech fluency parameters were measured (see Table 4-7) and then some derived fluency indices based on these measurements were calculated (see Table 4-8). Some aspects of these measures are explored more fully in Duff et al. (forthcoming), and this will be briefly touched on in the next chapter (see 5.3.3).

**Table 4-7: Observed fluency measurements derived from interview speech samples**

Measurement	Parameters / Notes
1. Total non-repeated Chinese syllables	<ul style="list-style-type: none"> <li>Total number of non-repeated Chinese syllables</li> <li>Syllables were discounted (considered 'repeated') if stumbled over (spoken more than once to no rhetorical effect) or used again to restart a sentence after a filled or unfilled pause (i.e. false starts or repetitions). (Hence non-repeated Chinese syllables should match the number of characters transcribed for that time period of participant speech.)</li> <li>Taken from the sections of the transcripts matching the audio samples</li> </ul>
2. Total number of syllables of English words	<ul style="list-style-type: none"> <li>Syllables of English words used when code-switching; not included in (1)</li> <li>Counted manually when listening to the audio samples</li> <li>Represented in transcript</li> </ul>
3. Total Chinese syllables of self-repairs	<ul style="list-style-type: none"> <li>Chinese syllables deemed not a part of (1)</li> <li>Counted manually when listening to the audio samples</li> <li>Not consistently represented in transcript<sup>30</sup></li> </ul>
4. Total filled pauses	<ul style="list-style-type: none"> <li>A filler pause was any audible self-repair that was not a semantically identifiable morpheme (e.g. um, ah, en)</li> <li>Measured using Audacity</li> <li>Total length of filler pauses was not measured</li> </ul>
5. Total unfilled pauses	<ul style="list-style-type: none"> <li>A pause was considered unfilled if silent, sentence-internal, and of duration 0.2 seconds or more</li> <li>Measured using Audacity</li> </ul>
6. Total length of unfilled pauses	<ul style="list-style-type: none"> <li>The sum of the individual lengths of all pauses from (5)</li> <li>Observed using Audacity</li> </ul>

---

<sup>30</sup> For the most part obvious repetition that was quickly corrected was removed from the transcripts, but some repetition due to unfilled pauses deemed to be 'pause for thought' did remain in the transcripts; see 3.3.1 for further discussion of transcription processes and consistency.

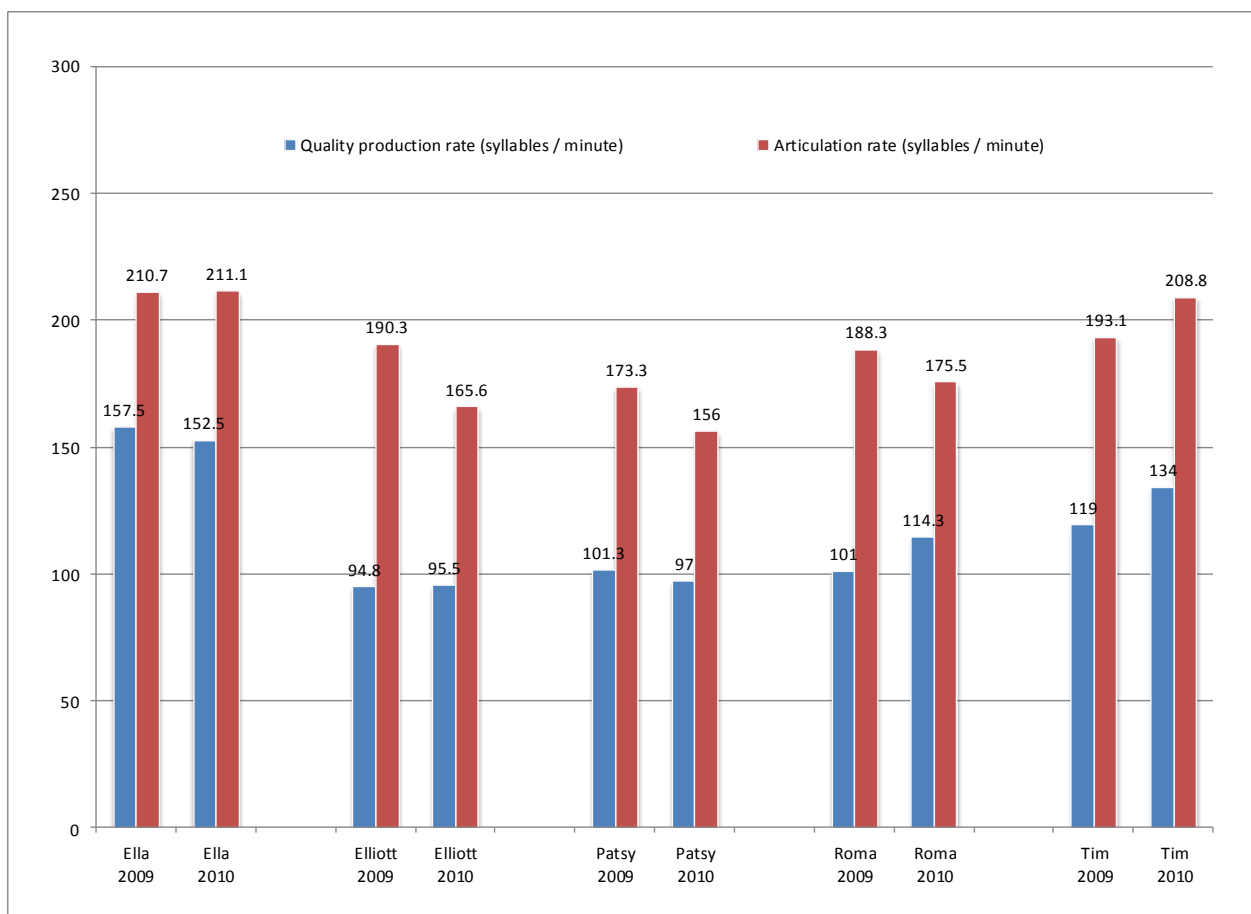


**Table 4-8: Calculated fluency indices derived from observed fluency measurements**

Calculated fluency indices	Parameters / Notes
7 Quality production rate	<ul style="list-style-type: none"> <li>Number of non-repeated syllables / minute</li> <li>(1) divided by total length (four minutes)</li> <li>In Duff et al. (forthcoming), called “oral fluency”</li> </ul>
8. Articulation rate	<ul style="list-style-type: none"> <li>Number of non-repeated syllables / (total length minus unfilled pauses)</li> <li>(1) divided by the difference of total length (four minutes) and (6)</li> <li>In Duff et al. (forthcoming), called “rate of speech”</li> </ul>
9. Mean unfilled pause length	<ul style="list-style-type: none"> <li>Average length of unfilled pauses in seconds</li> <li>(6) divided by (5)</li> </ul>
10. Mean length of run	<ul style="list-style-type: none"> <li>Average number of syllables produced before pausing</li> <li>(1) divided by (5)</li> </ul>
11. Mean length of utterance	<ul style="list-style-type: none"> <li>Average length in seconds of an utterance before pausing</li> <li>The difference of total time (four minutes) and (6), divided by (5)</li> </ul>
12. Self-repair rate	<ul style="list-style-type: none"> <li>Repairs per minute</li> <li>(3) divided by total length (four minutes)</li> </ul>
13. Ratio of pruned length to total length (RPL)	<ul style="list-style-type: none"> <li>The proportion of self-repairs to total speech</li> <li>(3) divided by the sum of (1) and (3)</li> </ul>
14. Disfluency	<ul style="list-style-type: none"> <li>Proportion of self-repairs, filled pauses, and English to total speech</li> <li>The sum of (2), (3), and (4) divided by the sum of (1), (2), (3), and (4)</li> </ul>

#### 4.4.1 Speech rate

The calculated speech rate measures (7) and (8) are presented in Figure 4-6.



**Figure 4-6: Speech rate fluency indices for each participant**

Quality production rate here shows how much intended, comprehensible speech was produced per minute, including pauses. Articulation rate, which discounts unfilled pauses, is a rough indication of how quickly a participant spoke when actually producing speech; it mitigates the effects of frequent and extended pausing for thought, focusing on how quickly a participant spoke once the words were decided on.

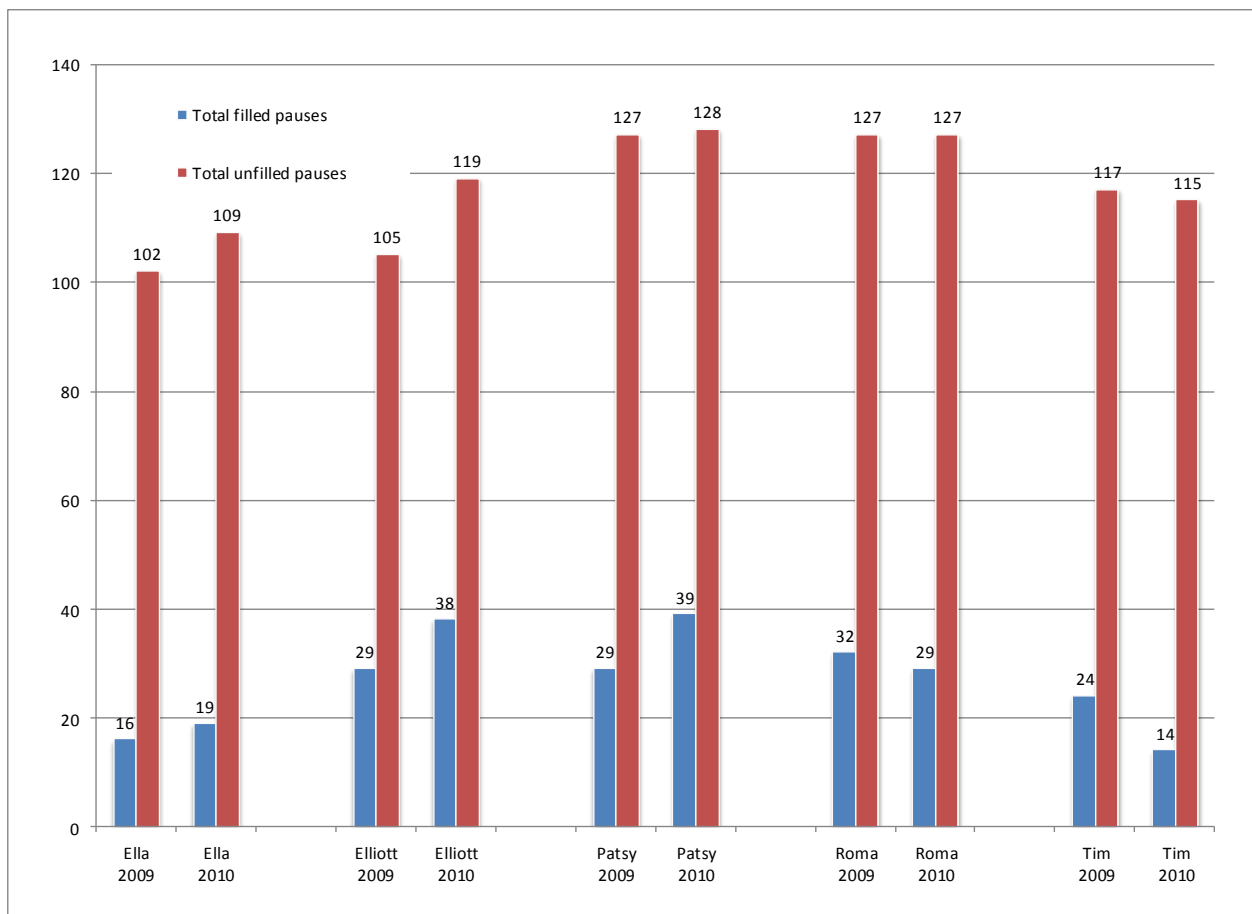
Ella, who self-assessed around the B2 range in oral proficiency, showed consistently high results, speaking quickly and with a high quality production rate, though the increase she noted from 2009 to 2010 (from B2- to B2+) did not appear evident in these measures, perhaps an indication of a sort of ‘ceiling effect’ indicating a non-linear relationship of these measures to overall proficiency. Tim’s increase did seem borne out by these measurements, and in fact despite self-assessing at A2 to B1- he exhibited greater fluency by these measures

than others who self-assessed globally at a higher level, namely Elliott and Roma. Elliott's scores were lowest for quality production rate, as he paused more than any other participant, and for longer; his articulation rate, however, was comparable to Roma's, showing a more comfortable B1-comparable rate of speech when actually producing speech.

Given that both Tim's and Elliott's scores were somewhat different than expected, and that some but not all predicted increases in proficiency were borne out by these indices, speech rate measures appear to be promising but only one factor among many in characterizing overall oral proficiency.

#### **4.4.2 Pauses**

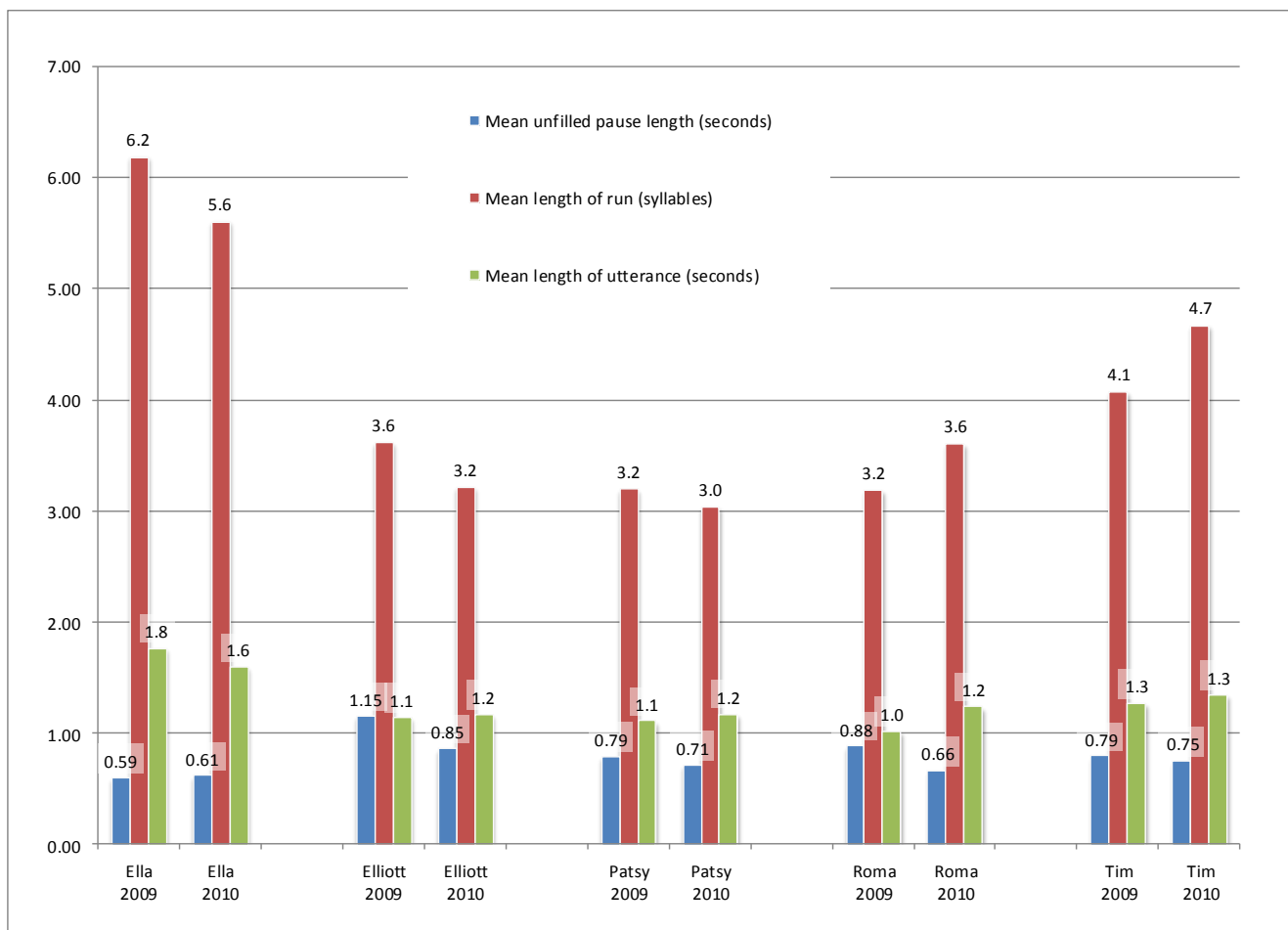
The pause indices are derived from measures (4), (5), (9), (10), and (11) in Table 4-7 and Table 4-8 above. Pause counts are shown in Figure 4-7; length indices are shown in Figure 4-8.



**Figure 4-7: Fluency indices: Pause counts**

We can see that once again, as for the speech rate measurements described in section 4.4.1, Ella's B2-level oral proficiency is evident in her pause counts, which are lower than those of the other participants. Roma and Patsy appeared to be consistent, and in fact Roma's slight decrease in filled pauses reinforces the gains she saw in quality production rate (see section 4.4.1).

Again, of the two participants who self-assessed at a higher oral proficiency level in 2010 compared to 2009, these measurements only show increased fluency for one, Tim; Ella paused slightly more often in 2010, though the differences for both Ella and Tim are slight, and no greater in magnitude than differences observed between the two interviews for the other participants – Elliott in particular saw a 30% increase in filled pauses and a 13% increase in unfilled pauses from 2009 to 2010, possibly indicating greater hesitancy or thought put into his interview responses.



**Figure 4-8: Fluency indices: Pause length and speech length**

Once again the pause length and speech length indices strengthen the case for Ella's relatively high oral fluency. Her mean unfilled pause length measures are low. Roma and Elliott show the greatest improvement in

mean unfilled pause length from 2009 to 2010. Ella and Tim both had relatively high scores for mean length of run, and again Tim showed improvement in this area from 2009 to 2010 (and the change was by the largest percentage margin of the group), while Ella did not. Finally, Ella's mean length of utterance was also clearly relatively high compared to the other participants.

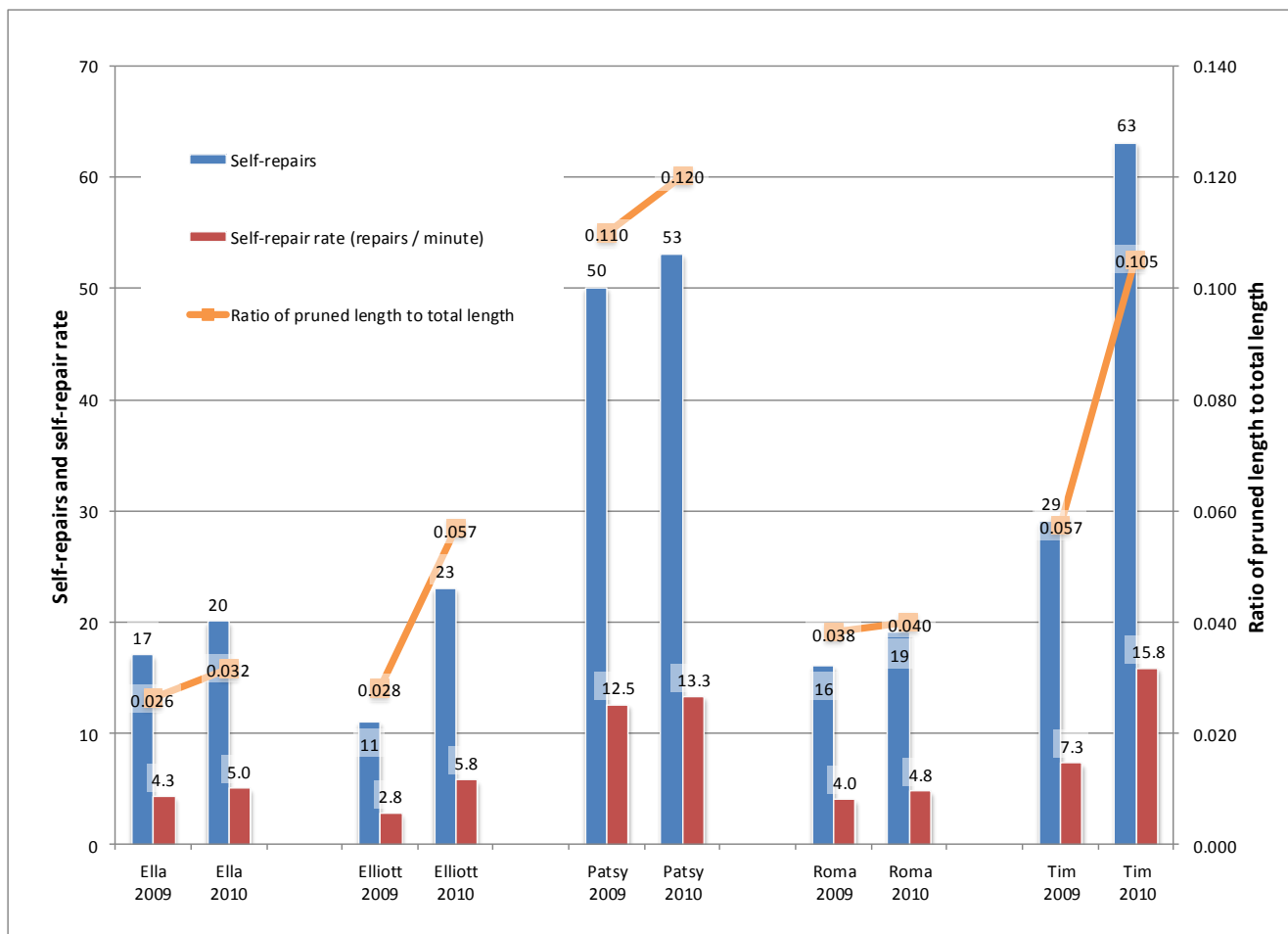
Of the CAL-related pause fluency measures calculated,<sup>31</sup> there does appear to be a correlation to self-assessed oral proficiency, though like the speech rate measures, there is some inconsistency to the expected gradation, and discernment does not seem to be nuanced enough to account for differences around the A2-to-B1 range.

#### **4.4.3 Self-repairs**

Measure (3) in Table 4-7 above can be used to calculate the repair fluency indices (12) and (13); the results for the interview data are shown in Figure 4-9.

---

<sup>31</sup> Other pause ratios discussed in CAL literature (see section 2.3.3) are different ways of representing the indices above: False pause rate is more useful than total false pauses when samples are not of the same length; phonation-to-time ratio is a calculation made with the same numbers as mean false pause length.



**Figure 4-9: Fluency indices: Repair fluency**

In 2009, Elliott showed the lowest rate of self-repair, but this is perhaps to be expected considering his relatively high pause counts and the fact that his mean unfilled pause length was longest in the group; it is possible that pausing for longer allowed him to think long enough that he made fewer utterances requiring self-repairs (see 2.2.4.2 for a discussion of self-repair measures in the context of accuracy analysis). In general, Ella, Roma, and Elliott showed low self-repair rates. Ratio of pruned length-to-total length (RPL) attempts to give an indication of how many syllables relative to total production were self-repairs, and so is independent of speech rate. Here we see that Tim and Patsy exhibit relatively high RPL measures, indicating that they had more self-repairs in their speech on average than the other participants; although this might hint at a preponderance of

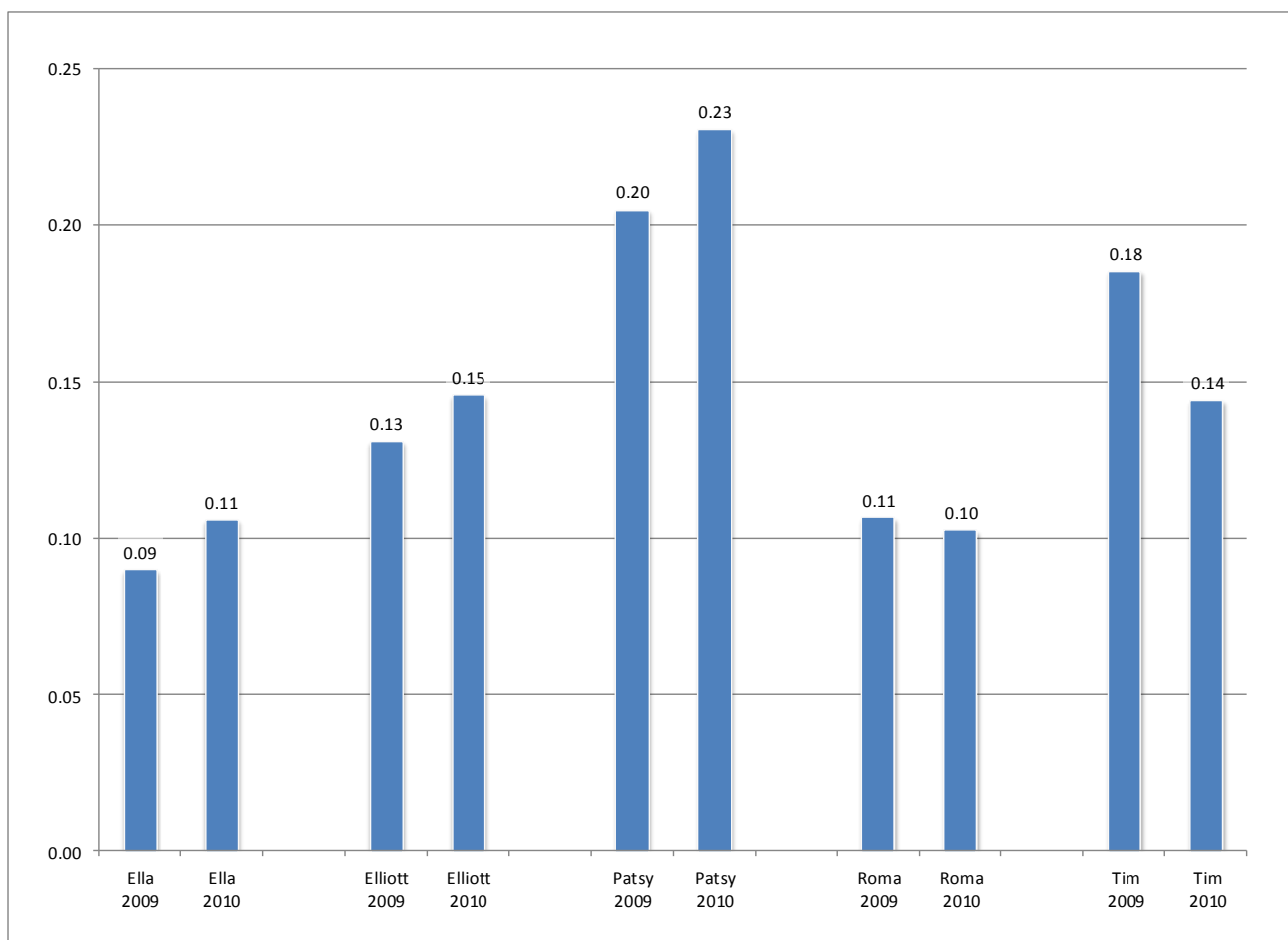
pausing and correction impeding the flow of speech, in Tim's case this did not seem to affect his quality production rate, though Patsy's was lower.

Self-repairs (in a sense, self-monitoring) certainly indicate fluency in that they give a picture of how often the speaker attempts to correct his or her utterances, which may impair the flow of speech, but without the context of why the repair was necessary or whether it was successful, it is again just one factor in a larger concept of fluency. On one hand, few repairs may mean few mistakes and confident speech; on the other, it may token poor linguistic awareness in a speaker who does make many mistakes, and in fact in this case the lack of repairs may detract from the comprehensibility of speech.

#### **4.4.4 Disfluency**

I also calculated disfluency; although the fluency calculations did not use units of words (see further discussion in section 2.3), I tried to follow the same principles laid out in section 2.3.3, using syllables as the unit of measure. The results of the calculations for the interview data (14) is presented in Figure 4-10.





**Figure 4-10: Fluency indices: Disfluency**

This fairly inclusive measure shows Ella and Roma as those least affected by disfluencies in speech, followed by Elliott, and lastly Tim and Patsy. These numbers follow the general expectations given by their respective CEFR self-assessments for oral proficiency, with the exception being Elliott, whose 2010 disfluency score is similar to Tim's; this is perhaps unsurprising, and corroborates well with the longer pauses Elliott took and relatively low quality production rate.

Disfluency appears to be a more robust measure than e.g. pause rate or RPL, which take into account only one aspect of non-intended utterances. Although disfluency doesn't incorporate pause length (or, indeed,

the effects of pauses or self-corrections on accuracy or comprehensibility), it does seem to provide a more holistic view of how each of its component factors detract from fluency.

## **4.5 Summary**

In this chapter we looked at the actual results of some of the theoretical indices and CAF techniques discussed in Chapter 2 for the interview data under study. It was possible to explore most of the aspects of complexity and fluency that have been explored in general CAF literature and CAL-specific CAF literature. In some cases, such as for lexical measures and fluency, I was able to apply the index concepts easily despite the typology of Chinese being different than many more commonly researched languages, and despite its shorter history of CAF study, which meant less robust reference materials. In other cases, such as characterizing syntactic complexity and syntactic variety, the difference in typology between languages in previous CAF studies and Chinese made for a greater challenge, as tools and methodologies for quickly and comprehensively analyzing syntactic properties are less developed for Chinese. In Chapter 5 I'll present the implications of the results laid out here, and propose future extensions of the work that has already been done.

## **Chapter 5. Discussion**

### **5.1 Introduction**

This chapter revisits observations from Chapters 2, 3, and 4 in an attempt to integrate past and present results and put forth suggestions and cautions for future work. First it summarizes from a more global perspective the various findings in the present study for each participant and explores how CAF indices can be used to show change in proficiency over time. It then discusses each aspect of CAF and attempts to make recommendations for best practices for each element, summing up previous CAL-context CAF studies, comparing them to the current study, and adding in a few more observations from recent CAF studies applied in other foreign language contexts. Finally, this chapter examines the implications of these recommendations and sketches out some possible extensions of CAF research in the CAL context.

### **5.2 Comprehensive profiles: Change over time**

We can examine the overall CAF profiles for each participant, and their change from 2009 to 2010, to see a broad view of language proficiency measures, as in Table 5-1 through Table 5-5. Before discussing the participants, it's important to note that the changes shown in green and red represent only the data, not the qualitative interpretation of the data; a 'positive change' in percentage may not always imply an increase in proficiency. For example, a positive change in MATTR is arguably a positive result in terms of proficiency, whereas the import of a positive change in the percent of words used in a given frequency band or HSK level is more ambiguous (analysis would depend on the distribution in the other bands as well), and a positive change in mean unfilled pause length might indicate an actual decrease in proficiency (in this case, longer pause lengths might mean less confidence and lower fluency).

The data in Table 5-1 through Table 5-5 are not easy to absorb in one glance, and more work on how to integrate the various scales in terms of their actual relation to proficiency is necessary. For this thesis we can attempt to characterize in broad strokes the changes observed for each participant:

- Ella's proficiency does appear to have increased from 2009 to 2010 in terms of complexity. Her MATTR increased, and the percentage of words used in high-frequency bands and low HSK levels decreased, while low-frequency words use and high-level HSK use increased. Her fluency appears to actually have fallen somewhat; though production rate and articulation rate remained nearly constant, self-correcting behaviour, pause time, and disfluency increased, while mean length of run and utterance decreased, perhaps showing that Ella focused more on form in her second interview.
- Elliott's results point to the ambiguity in numbers without context. MATTR did not rise by much, and though he used a higher percentage of low-frequency words in his second interview, the decrease in use of low-level HSK words translated mainly into increased use of words not on the HSK, which makes it difficult to know if those words were more sophisticated or simply difficult to match to an HSK level without the kind of detailed look at the data we saw in section 4.2.2. Elliott's fluency measures did not show consistent improvement either: his articulation rate slowed, and mean length of run shortened, but mean unfilled pause length also shortened, and mean length of utterance lengthened. His disfluency measure increased, but the higher rate of self-repair may have indicated a desire to produce more comprehensible and more accurate speech.
- Patsy's complexity measures improved: her MATTR increased, her use of low-frequency words dropped, and her use of words from HSK levels 4 and 5 rose dramatically. However, her fluency measures indicate that her speech was slower and more disfluent (with more self-corrections) in her second interview, perhaps an indication of trade-off between complexity and fluency.
- Roma's variety and sophistication of lexical expression dropped from 2009 to 2010 according to her fluency numbers: her MATTR shifted slightly downwards, and she used in general a higher percentage

of high-frequency words and low-level HSK words, though like Elliott her results are difficult to interpret because her use of words not within the LCMC-5000 or the HSK list increased as well. Roma's fluency, however, showed a proficiency increase from 2009 to 2010, as unfilled pause length dropped dramatically, quality production rate increased, and self-repairs went up without heavily affecting RPL. Perhaps Roma focused less on form in the second interview, which boosted her confidence in producing fluent speech.

- Tim's proficiency increased from 2009 to 2010, as his MATTR improved and he used more high-level HSK vocabulary (though his word frequency results were difficult to interpret at a glance). Most telling, his fluency increased dramatically according to the measures, and though he had a higher rate of self-repairs they appear to have been part of an overall greater awareness of the language, as his articulation and quality production rates improved significantly.

Taking an overall picture of the changes a participant's speech showed over time in such a way can be helpful in understanding or contextualizing qualitative observations and changes in life trajectory, such as validating the effects of Ella's and Tim's formal studies between the 2009 interview and the 2010 interview. They also may help point out in a quantitative way the effects of focus on one aspect of proficiency (e.g. complexity) on another aspect of performance (e.g. fluency).

**Table 5-1: Ella's CAF profile and changes over time**

<b><u>Complexity</u></b>	<b>Ella 2009</b>	<b>Ella 2010</b>	<b>Percent change</b>
MATTR	0.58	0.62	+8.0%
LCMC range 1-500	46.6%	42.6%	-8.7%
LCMC range 501-1000	15.7%	12.9%	-18.2%
LCMC range 1001-2000	14.0%	12.7%	-9.8%
LCMC range 2001-3000	4.2%	5.5%	+31.6%
LCMC range 3001-4000	6.5%	5.1%	-20.3%
LCMC range 4001-5000	2.2%	2.2%	-3.1%
>5000	10.7%	19.0%	+78.1%
Mean word frequency of unique words on LCMC 5000 list	927	961	+3.6%
Mean word frequency for all occurrences of words on LCMC 5000 list	160	169	+5.5%
HSK level 1	17.7%	17.2%	-2.6%
HSK level 2	14.3%	11.9%	-17.1%
HSK level 3	15.4%	15.0%	-2.6%
HSK level 4	14.6%	12.5%	-14.6%
HSK level 5	6.7%	7.7%	+14.6%
HSK level 6	2.0%	2.4%	+20.8%
Not in HSK	29.2%	33.3%	+13.9%
<b><u>Fluency</u></b>	<b>Ella 2009</b>	<b>Ella 2010</b>	<b>Percent change</b>
Quality production rate (syllables / minute)	157.5	152.5	-3.2%
Articulation rate (syllables / minute)	210.7	211.1	+0.2%
Mean false pause length (seconds)	0.59	0.61	+2.8%
Mean length of run (syllables)	6.2	5.6	-9.4%
Mean length of utterance (seconds)	1.8	1.6	-9.6%
Self-repair rate (repairs / minute)	4.3	5.0	+17.6%
Ratio of pruned length to total length	0.026	0.032	+20.8%
Disfluency	0.09	0.11	+17.8%

**Table 5-2: Elliott's CAF profile and changes over time**


























<b><u>Complexity</u></b>	<b>Elliott 2009</b>	<b>Elliott 2010</b>	<b>Percent change</b>
MATTR	0.61	0.63	+2.5%
LCMC range 1-500	38.7%	34.7%	-10.3%
LCMC range 501-1000	16.6%	14.9%	-10.3%
LCMC range 1001-2000	14.2%	13.9%	-2.5%
LCMC range 2001-3000	5.7%	5.9%	+4.7%
LCMC range 3001-4000	5.1%	6.1%	+19.2%
LCMC range 4001-5000	2.9%	1.9%	-36.3%
>5000	16.8%	22.7%	+35.1%
Mean word frequency of unique words on LCMC 5000 list	1013	1062	+4.9%
Mean word frequency for all occurrences of words on LCMC 5000 list	214	260	+21.3%
HSK level 1	11.5%	11.0%	-4.3%
HSK level 2	9.3%	10.2%	+9.1%
HSK level 3	13.3%	12.4%	-7.3%
HSK level 4	15.1%	11.5%	-24.0%
HSK level 5	11.1%	8.6%	-22.5%
HSK level 6	5.1%	4.7%	-7.3%
Not in HSK	34.5%	41.6%	+20.7%
<b><u>Fluency</u></b>	<b>Elliott 2009</b>	<b>Elliott 2010</b>	<b>Percent change</b>
Quality production rate (syllables / minute)	94.8	95.5	+0.7%
Articulation rate (syllables / minute)	190.3	165.6	-13.0%
Mean false pause length (seconds)	1.15	0.85	-25.6%
Mean length of run (syllables)	3.6	3.2	-11.1%
Mean length of utterance (seconds)	1.1	1.2	+2.2%
Self-repair rate (repairs / minute)	2.8	5.8	+109.1%
Ratio of pruned length to total length	0.028	0.057	+101.3%
Disfluency	0.13	0.15	+11.2%

**Table 5-3: Patsy's CAF profile and changes over time**

<b><u>Complexity</u></b>	<b>Patsy 2009</b>	<b>Patsy 2010</b>	<b>Percent change</b>
MATTR	0.51	0.55	+8.5%
LCMC range 1-500	46.7%	43.2%	-7.5%
LCMC range 501-1000	14.3%	13.1%	-8.3%
LCMC range 1001-2000	10.6%	12.1%	+14.3%
LCMC range 2001-3000	4.9%	4.5%	-6.7%
LCMC range 3001-4000	4.6%	5.3%	+15.7%
LCMC range 4001-5000	3.4%	3.0%	-11.9%
>5000	15.5%	18.7%	+20.8%
Mean word frequency of unique words on LCMC 5000 list	917	964	+5.2%
Mean word frequency for all occurrences of words on LCMC 5000 list	114	130	+13.2%
HSK level 1	23.5%	23.2%	-1.1%
HSK level 2	16.6%	15.7%	-5.8%
HSK level 3	15.5%	13.4%	-13.5%
HSK level 4	4.6%	7.3%	+59.7%
HSK level 5	1.1%	3.0%	+164.4%
HSK level 6	1.4%	1.3%	-11.9%
Not in HSK	37.2%	36.1%	-3.1%
<b><u>Fluency</u></b>	<b>Patsy 2009</b>	<b>Patsy 2010</b>	<b>Percent change</b>
Quality production rate (syllables / minute)	101.3	97	-4.2%
Articulation rate (syllables / minute)	173.3	156	-10.0%
Mean false pause length (seconds)	0.79	0.71	-9.7%
Mean length of run (syllables)	3.2	3.0	-4.9%
Mean length of utterance (seconds)	1.1	1.2	+5.6%
Self-repair rate (repairs / minute)	12.5	13.3	+6.0%
Ratio of pruned length to total length	0.110	0.120	+9.4%
Disfluency	0.20	0.23	+12.6%



**Table 5-4: Roma's CAF profile and changes over time**

<b><u>Complexity</u></b>	<b>Roma 2009</b>	<b>Roma 2010</b>	<b>Percent change</b>
MATTR	0.55	0.53	 -3.2%
LCMC range 1-500	48.8%	44.6%	 -8.5%
LCMC range 501-1000	12.6%	13.1%	 +3.7%
LCMC range 1001-2000	9.9%	11.6%	 +18.1%
LCMC range 2001-3000	6.6%	6.5%	 -0.6%
LCMC range 3001-4000	6.3%	4.5%	 -27.9%
LCMC range 4001-5000	1.9%	1.4%	 -25.9%
>5000	14.0%	18.2%	 +30.1%
Mean word frequency of unique words on LCMC 5000 list	923	896	 -2.9%
Mean word frequency for all occurrences of words on LCMC 5000 list	143	146	 +2.3%
HSK level 1	19.7%	21.6%	 +9.5%
HSK level 2	14.8%	15.9%	 +7.5%
HSK level 3	16.4%	13.4%	 -18.8%
HSK level 4	12.3%	9.4%	 -24.0%
HSK level 5	5.2%	4.8%	 -7.2%
HSK level 6	2.2%	2.3%	 +3.7%
Not in HSK	29.3%	32.7%	 +11.4%
<b><u>Fluency</u></b>	<b>Roma 2009</b>	<b>Roma 2010</b>	<b>Percent change</b>
Quality production rate (syllables / minute)	101	114.3	 +13.2%
Articulation rate (syllables / minute)	188.3	175.5	 -6.8%
Mean false pause length (seconds)	0.88	0.66	 -24.7%
Mean length of run (syllables)	3.2	3.6	 +13.1%
Mean length of utterance (seconds)	1.0	1.2	 +21.4%
Self-repair rate (repairs / minute)	4.0	4.8	 +18.8%
Ratio of pruned length to total length	0.038	0.040	 +4.8%
Disfluency	0.11	0.10	 -3.8%

**Table 5-5: Tim's CAF profile and changes over time**

<b><u>Complexity</u></b>	<b>Tim 2009</b>	<b>Tim 2010</b>	<b>Percent change</b>
MATTR	0.50	0.53	+7.0%
LCMC range 1-500	42.7%	43.0%	+0.8%
LCMC range 501-1000	13.2%	14.5%	+10.1%
LCMC range 1001-2000	13.8%	11.0%	-19.7%
LCMC range 2001-3000	6.5%	6.7%	+3.5%
LCMC range 3001-4000	5.9%	4.9%	-16.2%
LCMC range 4001-5000	3.4%	2.0%	-39.6%
>5000	14.6%	17.7%	+21.4%
Mean word frequency of unique words on LCMC 5000 list	1033	920	-11.0%
Mean word frequency for all occurrences of words on LCMC 5000 list	127	140	+10.1%
HSK level 1	21.1%	20.3%	-3.4%
HSK level 2	14.9%	13.1%	-12.1%
HSK level 3	16.3%	14.8%	-9.0%
HSK level 4	11.0%	9.0%	-17.7%
HSK level 5	3.4%	3.8%	+12.1%
HSK level 6	1.4%	2.3%	+65.6%
Not in HSK	32.0%	36.6%	+14.4%
<b><u>Fluency</u></b>	<b>Tim 2009</b>	<b>Tim 2010</b>	<b>Percent change</b>
Quality production rate (syllables / minute)	119	134	+12.6%
Articulation rate (syllables / minute)	193.1	208.8	+8.1%
Mean false pause length (seconds)	0.79	0.75	-5.0%
Mean length of run (syllables)	4.1	4.7	+14.6%
Mean length of utterance (seconds)	1.3	1.3	+5.9%
Self-repair rate (repairs / minute)	7.3	15.8	+117.2%
Ratio of pruned length to total length	0.057	0.105	+83.2%
Disfluency	0.18	0.14	-22.3%

## 5.3 Reflections on CAF analysis

This section revisits and tests the observations made in previous CAL-context CAF studies (as outlined in Chapter 2). As explained further in section 5.4, the data and analysis in this thesis is exploratory and thus in many cases agreement with previous results or the advancement of new hypotheses is limited by the amount of data and the fact that I used descriptive rather than inferential statistics to indicate observed differences.

Based on previous studies and any corroborating or challenging evidence from the current study, the following measures are deemed most promising for investigating links between proficiency and CAF indices (detailed discussion follows, separated into CAF sub-sections), taking into account ease of operationalization and automation:

- **Complexity:** MATTR, HSK and word frequency measures (with caveats, see 5.3.1.2 below)
- **Accuracy:** None (see 5.3.2 and 5.4 for further discussion of this limitation)
- **Fluency:** Speech rate (quality production rate, articulation), pauses (length, frequency), disfluency

Of the other indices explored or suggested in this thesis, many were too difficult to operationalize, and associating them clearly with proficiency would require more data or time to develop more robust indices.

### 5.3.1 Complexity

#### 5.3.1.1 *Lexical variety*

Yuan (2010) and the studies cited in Yuan (2009) drew the conclusion that TTR and MSTTR were of no significance to oral proficiency; Yuan (2009) did see a difference in lexical use between the NS and NNS used as examples in that study, but was skeptical of existing TTR measures. Iwashita et al.'s (2008) study of English for Academic Purposes test-takers found that vocabulary (variety of types, and token amounts per unit time, viewed as a fluency measure in the current study), MLU (again, viewed as a fluency measure in this thesis), and

verb phrase complexity, in descending order of strength of effect, were complexity measures related to spoken proficiency; however, vocabulary showed the only non-marginal effect of the three. Given the inconclusiveness of the studies to date on the relation between lexical variety measures and human-judged or standardized test measures, it is still important to take a measurement. Indeed, some prominent theorists in the field of CAF studies even believe that lexis is worth dissociating from complexity for independent research (Skehan, 2009). In any case, lexical variety measures may be able to show nuances in language proficiency gains or task difficulties that general scales like the CEFR are not suited to.

Of the measures examined in this study, MATTR seems promising. At this stage, in terms of the maturity of CAF studies in the CAL context, I believe TTR is still important to calculate, as its operationalization is simple and comprehensible, albeit flawed, and we see from the results of TTR calculations here that it does seem to conform to self-assessments. MATTR was calculated instead of other measures discussed in literature, namely MSTTR and  $D$ , due to the combination of algorithmic clarity<sup>32</sup> and text-size agnosticism ( $D$  may not be consistent across different-sized data sets, and MSTTR has truncation issues for data sets whose length are not exact multiples of the block size). It showed differences among participants, and also showed NS and NNS differences, though with just one exploratory NS excerpt this is tentative at best. Note that the NS excerpt used just 650 tokens (310 unique), but given the tests done for MATTR with windows of different sizes (see section 4.2.1), MATTR 100 should still work well on a data sample of that size. Note as well that while the NS excerpt was from a discussion, approximating the conditions for participants in this study, it was a high-stakes discussion on a nationally syndicated television show; further, as mayor of a Chinese city, the NS was probably a well-studied orator. The transcription available may also have corrected disfluency or errors in the NS speech to a degree. This sample NS comparison attempts to show differences at high proficiency levels, as the NS is an

---

<sup>32</sup> A module for calculating  $D$ ,  $\text{vocD}$ , is available for CHAT-formatted data part of CLAN (Malvern & Richards, 2002), but CHAT formatting is labour intensive and geared towards child speech.

educated adult, but this study has no comparison to native-speakers at different proficiency or developmental levels (e.g. children) to determine how greatly MATTR differentiates between low and high proficiencies for Chinese speech. In summary, the NS comparison given here is merely provided as an indication of the methodology; future work should more carefully match NS reference data in context (e.g. similar informal dialogue on similar questions, such as the picture description task, preferably with the same interviewer in the same context).

In terms of window size, one limitation of this study is that MATTR higher than 100 was not explored; it is possible that MATTR 100 is not in fact optimal for the data used herein. Among the window sizes examined, MATTR 100 was most appropriate for the data in this study, possibly due to the conversational nature of the data: conversation involves turn-taking, which limits the length of a single thought, and lends itself to an evolving stream of topics, but the thread of conversation, and in this particular case the fact that participants were mainly encouraged to speak at length about their experiences and thoughts, means that a certain amount of adherence to topic over time is still important. Perhaps for other studies it might be justified to explore whether a window size (ratio) relative to total text size is more appropriate, or whichever of a preliminary round of MATTR window variance shows greatest discriminatory power. One might also try to calculate or associate a particular MATTR range with a variety of genres of speech or text to identify bands of proficiency, if sufficiently discriminatory.

Future studies might wish to explore salient collocations (e.g. collocations or word bundles with English or CAL-specific characteristics) and lexical frequency analysis (e.g. words used more or less frequently than native speakers) as an index of lexical variety.

#### **5.3.1.2 *Lexical sophistication***

Past CAL-context research on this construct does indicate it is worth continued investigation: Yuan (2009) found that NS speech was more lexically sophisticated than NNS speech, and Yuan (2010) observed an

effect of focus-on-form on lexical sophistication. Yuan (2009) also noted that in one earlier investigation, study abroad did affect this measure.

#### *5.3.1.2.1 HSK and LCMC analysis*

In our data, HSK and word frequency measures appeared uncomplicated at first glance, but as shown in 4.2.2 in reality these measures can be surprisingly difficult to operationalize and interpret. The methodology behind the formulation of the HSK word list is opaque, and the results of the method used in this thesis show that comparisons between the raw output of a text segmenter such as ICTCLAS and the HSK list are confusing at best, misleading at worst. It is perhaps possible that it is easier to draw conclusions based on the distribution of words within the HSK list levels than from the number outside the HSK lists (that is, a high proportion of high-level HSK vocabulary may indicate higher proficiency, whereas a high proportion of vocabulary from outside the bounds of the HSK list may not, at least according to the methods used in this thesis); for instance, comparisons of Patsy and Tim, who self-assessed at the same CEFR level, seem to show that intra-level differences in lexical sophistication are visible using these techniques. However, given that a considerable proportion of words from each participant's interview data did not appear on the HSK list, yet showed characteristics that suggested they should be assignable to HSK levels (again, see 4.2.2), this is still at best a rough approximation. Furthermore, all participants in the current study used vocabulary from across the HSK levels, and no easy correlation between a given self-assessment and use of vocabulary from a single HSK level matches with the general level alignments proposed by Hanban (see section 2.2.1).

Interestingly, Chang (2011) highlighted the wide discrepancy between the lexical command expected of learners at HSK and TOCFL levels that are otherwise equivalent according to their claimed CEFR alignments; for instance, recommended vocabulary size for B2 proficiency was 5,000 words according to SC-TOP and just 1,200

words according to Hanban. The differences may be related to the relatively sparse HSK word list; for instance, the TOCFL vocabulary list<sup>33</sup> includes many word forms multiple times, each time associated with a different part of speech. Such a sparse HSK word list implies that comparing only exact matches between participant speech and the reference list may be inadequate for representing lexical sophistication in a manner consistent with Hanban's intentions when creating the list. Future studies might examine whether the TOCFL list is in a form more appropriate to this kind of analysis.

As for the LCMC, it's worth noting again that it is not an oral corpus, so the frequency data derived from it may not map well onto oral genres like conversation and interviews. In fact, the HSK list used in this thesis is also not specific to oral data; its use here as reference data for oral production is exploratory and intended to illustrate a possible methodology for examining lexical sophistication, not an optimal one.<sup>34</sup>

These findings can support the acknowledgement that no single word list can satisfy all research purposes, and the recognition that uncritical adoption of this lexical sophistication technique, particularly without indication of alignment choices, comparison methodology, or matching judgments, is a questionable pursuit. Lexical sophistication is still an interesting aspect of linguistic complexity, however, and since comparisons of word use to reference lists present such an intuitive measure for learners and educators, and previous research in the CAL context (see 2.3.1) supported its usefulness, it is worth further investigation.

Future work on refining this concept in the CAL context might look at devising a method of determining if a raw word not falling within the range of a given reference list (e.g. the HSK vocabulary list) is a compound made up of, or otherwise recognizably attributable to, a word on that reference list; care must be taken to develop systematic interpretations of parts of speech, and extrapolations of what the reference list selections

---

<sup>33</sup> The TOCFL vocabulary list was retrieved from <http://www.tw.org/tocfl/> on May 7, 2012.

<sup>34</sup> Guidelines and reference data were made available for the HSK oral test in 2011, but were not available for the analysis in this thesis, and I could find no research that had explicitly made use of these resources at the time of writing. For more information see [http://product.dangdang.com/product.aspx?product\\_id=21083125](http://product.dangdang.com/product.aspx?product_id=21083125) (accessed May 7, 2012).

imply learners should know for a given level (such as expectations of combinatorial possibilities, e.g. with prefixes, suffixes, and bound forms; also morphosyntactic sophistication in the use of basic morphemes in complex ways, e.g. polysemic considerations, discussed next). Further, oral corpora should be used for frequency references in oral contexts, preferably segmented by the same parsing tools being used for the analysis of individual performances. Finally, for data such as the interview data used in this thesis, oral proficiency vocabulary lists should be used in place of general pedagogic vocabulary lists in future refinement of this type of index.<sup>35</sup>

#### 5.3.1.2.2 *Polysemic analysis*

The specific polysemy measures examined here (focusing on morphemes associated with complement structures, a form of verb phrase complexity) were perhaps too crude or too few to prove correlation to proficiency, but they appear promising (there is discernment, and greater variety does seem to correspond with high self-assessment). In the future, expanding the range of polysemous morphemes under investigation, and coding them according to authoritative dictionaries or grammars is recommended. Depending on context, we could still imagine a native speaker using only a few of the various meanings in a given instance of speech, so it's unclear whether or not there would be a leveling effect after the intermediate level, or even indeed a drop-off of variety at native-speaker-like levels similar to Jin (2007)'s observations on T-unit measures.

#### 5.3.1.2.3 *Word frequency analysis*

When examining word frequency measures, why not also look at word frequency range (high-low)? It seems that for long strings of speech, the fact that a given word is 'very low frequency' compared to others

---

<sup>35</sup> Although Chinese national standards for oral CFL proficiency were developed for introduction in 2011, the reference materials were not available for the current study. For further information, please see: 外国人考汉语有了口语国家等级标准 [Foreigners taking Chinese tests now have national oral level standards] (2010). 新华网天津频道 [Xinhua Net Tianjin]. Retrieved May 6, 2012, from [http://www.tj.xinhuanet.com/2010-12/12/content\\_21608845.htm](http://www.tj.xinhuanet.com/2010-12/12/content_21608845.htm)



may have a great deal to do with genre and interview content, as opposed to word knowledge. We need large speech samples and averages to really see that a speaker uses low-frequency words often (statistically speaking), and of course the qualitative decision of ‘accuracy of communication’ is also important when dealing with low-frequency words.

### **5.3.1.3 Syntactic complexity**

According to Iwashita et al. (2008), in the context of oral proficiency, “studies... appear to show that grammatical accuracy is the principal determining factor for raters assigning a global score” (p.27). If this is true, the complexity of the grammar attempted is an important context in which to judge proficiency; adventurous learners who attempt a variety of complex utterances may exhibit a lower rate of accuracy than conservative learners who produce simpler utterances, so complex sentences may not be a strong indicator of overall proficiency.

This thesis did not attempt to calculate any measures for syntactic complexity, but there is discussion of one possible methodology using tools in the public domain in section 4.2.3, a methodology which supports some of the recommendations in previous studies (Jin, 2007; Yuan, 2009) to focus on topic-comment constructions and zero elements, for example. Future studies could make use of the rich output from segmenting libraries to examine POS variety, parse tree lengths, and different branching (subordination) types. Development of scripts to examine output of these libraries was deemed beyond the scope of this thesis due to time constraints and the poor-quality output of the freely available POS taggers examined.

Note that the publicly available libraries mentioned in this thesis are not state-of-the-art; a good deal of computational linguistic resources are expended in the service of modern search and retrieval technologies, and frequently tied to patented or proprietary private-sector research.<sup>36</sup>

I think that, ultimately, sacrifice of high accuracy from inter-rater-validated hand coding for speed and recognized, internationally criticized and refined standards for metadata ontologies is worthwhile, especially given the fact that linguistic software is getting more accurate every year, and better training data is validated for use in testing it. Thus, an important aspect of future work in this area will be operationalizing topic-comment recognition, covert conjunction coding, and other such methodologies for use with linguistic libraries.

Finally, evidence from work by Magnan (1988) comparing OPI scores with grammatical errors, found that “at higher levels, learners attempt more complex grammatical notions, and consequently make more errors” (as described in Iwashita et al., 2008, p.26). This has implications for can-do statements and methodologies like those used in this thesis, which reward attempts at complexity without penalizing for errors; there may still be a case for arguing that higher complexity equates to higher proficiency even if the trade-off is a higher degree of inaccuracy in production.

#### 5.3.1.4 *Syntactic variety*

This thesis looked at just a very few grammatical features to see if their presence or absence revealed anything about the participants’ proficiency levels. There does seem to be correspondence between a variety of syntactic uses of aspectual morphemes and proficiency, and though *ba*- and *bei*-constructions are considered basic by most CAL textbooks, the frequency of their presence also appeared to be related to self-assessed

---

<sup>36</sup> Even in the public domain, the resource-intensive nature of developing linguistic software means the latest technology is often available only for a fee. For instance, the Penn Chinese Tree Bank (CTB) 3.0 is available for free, but as of March 2012, the newest edition (7.0) is only available via Penn Linguistic Data Consortium for a fee. (The Treebank helps the statistical parser make good decisions; in the case of the Stanford Parser, as used in this thesis, it appears to use training data from the pre-release version of the 6<sup>th</sup> edition of the CTB.)

proficiency. Studies with other languages have examined a wider range of syntax; for example, Taguchi (2008) developed a measure based on examining a range of ‘chunks’ (semi-fixed morphemic patterns with grammatical salience, e.g. noun + *de*, noun + *ni*, clause + *kara*) to show grammatical variety for elementary learners of Japanese as a Foreign Language.

Building on solid research of individual aspects of Chinese grammar acquisition already established in the CAL context to expand the range of interesting grammatical particles and patterns surveyed would be a good next step for CAF research in the CAL context.<sup>37</sup> Examples of structures to investigate could include chained verb constructions, particles such as 的, 得, and 地, aspectual structures, complement structures, non-interrogative question words, classifier use, measure words, prefix and suffix noun modifiers, and so forth.<sup>38</sup>

### 5.3.2 Accuracy

Previous studies on accuracy measures have tried various operationalization techniques, including error frequency (previous studies have not led to strong conclusions), self-repair behaviour (again, tentative or inconclusive results), error type, and error gravity (these latter have not been operationalized consistently up to this point). The current study does not attempt any measures of accuracy, outside of looking for the presence of accurate use of words or grammar as part of complexity indices; due to the guiding principles of the thesis and the arrangement under which the data was made available to me, indications of how often or how severely participants used erroneous language are not included in the analysis.

---

<sup>37</sup> To some extent, since Chinese word choice affects grammatical function (instead of verb inflection in e.g. English), a shadow of syntactic variety will be notable as lexical variety, but evidence of function word variety is likely to be buried by the much-larger set of substance words (nouns, adjectives, verbs).

<sup>38</sup> A good place to start in determining which patterns and constructions would be interesting to investigate would be review articles of SLA research in the CAL context, such as Biq et al. (1996), Cui (2005), and Cao (2009). More specific examples include Shi (1998) and Ke (2005) for patterns of acquisition, Wen (2006) and Jiang (2009) for word order acquisition, and Xiao (2004, 2011) for discourse features. There are a plethora of studies published in China and internationally on CAL use of specific lexical or grammatical features as well, such as suffixes (Chen, 2009), directional complements (Wu, 2011), and aspectual meanings of negation markers (Yan, 2011), just to name a few. (Many of these studies could be used as a basis for local accuracy measurements as well.)

Moreover, as mentioned briefly in 4.3, accuracy measurements are not easily automated, as text segmenting libraries assume input is well-formed. To cut down on the raw amount of time and effort required for coding language for accuracy, however, some help can be derived by a human judge from tools such as word processors or concordancers to quickly show word sets and their contexts. Another way to speed up these measurements would be to pick an ‘indicative’ or ‘typical’ section of speech/text, e.g. several minutes of randomly sampled speech, and use that as a local approximation to a global accuracy measure.<sup>39</sup>

As for syntactic variety, discussed above, accuracy measurements could be drawn from considerable resources expended to date in CAL error analysis studies. One way of looking at this could be based on contrastive language analysis, e.g. modal verbs (e.g. use of 会 ‘can do’, 可以 ‘can do’, and 能 ‘can do’ as a many-to-one mapping to English ‘can/may’ and with overlapping usages of ‘will vs. is able to’), pronoun use (e.g. Mandarin tendency to omit pronouns or make use of topic-chaining vs. English requirement of pronouns), differences between use of English verb ‘to be’ and Chinese 是 (copula), etc.

Finally, given that Chinese is a tonal language, an investigation of tonal accuracy indices as an aspect of overall proficiency might be interesting. This technique was employed in Guo’s (2007) study, and research has been done in this area for non-tonal languages that could be studied and adapted or extended; an in-depth example of exploring phonological accuracy (pronunciation, intonation, and rhythm; in this case for L2 English) can be found in Iwashita (2008). It bears mentioning, however, that different Chinese communities have different standards for tone production and vowel quality, such that a Mandarin speaker from Taipei might tend to

---

<sup>39</sup> If accuracy is being assessed based on a transcript, it is important to note that judgments made during the transcription process may obscure or change the quality or quantity of errors (a ‘smoothing’ of the text, as it were), depending on the aims of the transcription. For instance, in this study repetitious syllables were not recorded in the transcripts, which would lead accuracy researchers astray if they had access only to the transcripts but wished to examine errors and repairs.

pronounce some words differently from a Beijing speaker; the question of what a researcher might determine to be ‘correct’ is outside the scope of this thesis but worth careful justification.

In further CAF studies (not solely in the CAL context) it may be worth investigating whether or not there are indications of significant correlation between accuracy and complexity or fluency measures. If so, it might be proposed that complexity and fluency measures, easier to automate, could be used as approximations for accuracy in certain contexts.

### **5.3.3 Fluency**

The fluency measures calculated in this thesis appear to have two main weaknesses. The first is that fluency measures are difficult to contextualize without reference to the specifics of the actual situation the language production occurred in, e.g. the pragmatic context of pauses and repetitions. Duff et al. (forthcoming) provides a more detailed exploration of the specific factors that have local effect on fluency measures, including concrete examples of, for example, the conflation of pausing for thought (independent of proficiency) and pausing to find the appropriate word (related to proficiency). The second major issue is that without the participants’ L1 fluency baseline data when discussing comparable topics, or for that matter NS baseline data discussing comparable topics, it is impossible to know what is ‘normal’ for speech rate or disfluency.

In general, fluency measures may be harder in the foreseeable future to completely automate, as (similar to phonological accuracy measures) they require audio processing capabilities. That is, aligning a transcript to a piece of audio and calculating pause lengths, turns, and so forth, is still a manual process. With better speech-to-text engines, and some scripting of audio processing tools, we may begin to see such issues tackled. For instance, speech-to-text conversions by necessity align words in transcripts with their exact locations in an audio file, thus calculations such as quality production rate and articulation rate (i.e. speech rate) can be made automatically, particularly if the speech-to-text functionality is sophisticated enough to mark and account for self-repairs and different pause types. Voice recognition in speech-to-text software is necessary for examining

speech in dialogue format, not just for turn-taking, but also to recognize which pauses are inter-sentential, which lengths of speech are the subject of fluency analysis (or are to be tagged in the transcript for further complexity or accuracy analysis), and so on. Pause recognition and length calculation are in principle not difficult to perform using computer scripting, but recognizing context (mainly who is talking) and pause type (filled pauses are difficult to detect without a sophisticated speech-to-text engine) are still challenges for automation.

An example of pushing into the territory of automation is Ginther et al. (2010); the paper analyzed Oral English Proficiency Test (OEPT) answers given by learners of English as a Foreign Language using automated speech evaluation (ASE) techniques (PRAAT, a custom Python program, and SAS). They compared the measures taken to holistic human ratings, learners' language backgrounds, and Oral English Proficiency Test (OEPT) scores and found that there was correlation between OEPT scores and measures of speech rate, mean length of run, and pause length and frequency, however these indices were not discriminatory enough to distinguish different levels of OEPT scores, and were thus deemed to be just part of the picture of oral proficiency and not perfect ways to boil down proficiency into a single set of measurements.

Note that in terms of aligning measures against overall competency, it's quite possible that slow and measured, but accurate, complex, and varied speech might be considered just as competent as quick, simple, direct speech by interlocutors; in different situations, different approaches to communication are also going to be important (an impassioned, reasoned speech might be slower and more complex; a conversation about the day might be rapid-fire but repetitive in terms of lexical variety). Thus, more studies of the type described in Iwashita et al. (2008), which examine what measures and indices actually correlate with native-speaker ratings in different situations, are going to be important in the field.

Past CAL-context CAF studies have not named any one fluency result as a clear-and-away winner, as they have examined these measures for different reasons (fluency, task performance, NS-NNS differences, etc.), but there are good indications that speech rate measures (as affected by focus on form, anxiety, and NS-NNS division) and breakdown fluency (e.g. unfilled pause rate and mean unfilled pause length) may be good leads.

Iwashita et al. (2008)'s study confirms this, along with Ginther et al. (2008)'s observations that speech rate, unfilled pauses, and pause time were indicative of proficiency; in short, "higher-level learners spoke faster with less pausing, and fewer unfilled pauses" (p.41). Exploratory results from this thesis do seem to correspond (though speech rate was more varied for this data set), and additionally disfluency was found to be a helpful measure.

## 5.4 Limitations

This study was intended to be summative, in the sense of reviewing the extent of CAF oral proficiency research in the CAL context, and also exploratory, by examining various classic and novel CAF measures for the transcribed speech of a small set of subjects. This section aims to recognize the various limitations of the methodology and source data.

**Statistical analysis:** One significant limitation of this study is that rigorous inferential statistical analysis was not used, in part because of the small sample size. For example, when determining the spread of high- vs. low-frequency words used by each participant, such analyses would be important for determining if there is indeed enough data (raw speech) for the results to be considered representative of the true usage spread over large amounts of speech. In cases where small fluctuations in observed values produce dramatic changes in the derived results (e.g. measures of disfluency based in part on a few counts of repairs or pauses) statistical analysis would help determine whether the derived value was a valid indication relative to the amount of measurements. This exploratory study could be made more robust by including common statistical analyses, e.g. inter-rater agreement on transcription choices, reliability checks to validate segmenter output, and ANOVA calculations for fluency calculations, analysis of polysemous word use, and grammatical construction use (see for example Iwashita et al. (2008)).

**Working with oral data; transcription choices:** Transforming audio (and indeed, visual, via gesture) source data into data that can be acted on by existing computational linguistic software packages is also an in-

herently imperfect process. Transcription choices, e.g. what to include and how to format the oral data, can be very difficult, and also inconsistent: formatting numbers in Roman or Chinese numerals, consistency of presentation of certain words or sounds such as interjections, standards for representing false starts and repetition, judgements surrounding unclear phonetics or which of several possible homophones was intended, and in the case of our transcription process, manual scanning of transcribed data to remove repetition in preparation for segmenting; these are just some examples of where uncertainty and inconsistency can affect outcomes. A concrete example is that although we chose to remove syllable and word repetition from our transcripts, we retained incorrect syllables alongside repairs, which can affect parsing, as statistical parsers such as ICTCLAS start from the assumption that a sentence is perfectly formed, and thus try to incorporate these errors into a working model of a sentence, without grasping their irrelevance to the substance of an utterance.<sup>40</sup>

**Parser accuracy:** Another area where statistical analysis is important, but outside the scope of this thesis, is in the accuracy<sup>41</sup> of the segmenting algorithms used by ICTCLAS and the Stanford Parser as used against corpora of conversational Chinese; the statistical success of the segmenting algorithms should be taken into account as an inherent, base element of uncertainty that feeds into all CAF calculations, as they assume perfect input. In light of the fact that inaccurate choices cannot be eliminated completely, I selected ICTCLAS for transcript segmentation based on its impressive accuracy statistics, its long history, its free availability, its speed and ease of use, and also on preliminary tests of ICTCLAS and the Stanford Parser using a subset of the interview, which showed ICTCLAS made better choices for our data set.

---

<sup>40</sup> One counterexample to our decision is Collentine (2004), who chose to only transcribe repairs (and not ‘incorrect words’) when looking at Spanish oral data; this might make parsing and lexical statistics slightly more accurate with respect to the intended meaning instead of the actual words spoken.

<sup>41</sup> Segmenting relies on a combination of statistical analysis and grammar-tree analysis, depending on the parser, and errors when evaluating a string of Chinese morphemes can lead to incorrect POS tagging or even incorrect or inconsistent identification of words from strings of syllables.



**Reference data:** Expanding on this theme, one of the fundamental methodological decisions in this thesis was to attempt to let automation and computers lighten the judgment and classification load on humans, and all automated processes are subject to error as the models used are often generalized approximations of a complex, subtle, and still not fully understood human heuristic for determining what is ‘right’ or ‘valid’ in language use. Models can be tested and tuned by seeing how their application compares against a set of reference data that has been hand-coded to some acceptable level of agreement. For this to truly lend legitimacy, the reference data must be considered to belong to an acceptably similar ‘genre’ to the text to be automatically analyzed by the model, which is a much-harder-to-quantify comparison, therefore standardizing or getting up to an objective level of agreement about such textual genre equivalencies might be tough.

Oral reference data is one of the limitations of the analysis in this thesis. For example, whether ICTCLAS or the Stanford Parser have been validated against oral data is unclear, so their accuracy for these types of data sets may be lower than the reported validation accuracy, which is likely based on written reference data. This lack of oral reference data muddies other analytical waters as well: I could not find any NS or NNS corpora with similar contexts (oral, interview, low-stakes, mutual familiarity) in the public domain, which means that testing our CAF measures against pre-tagged and validated NS or NNS data was impossible. The lack of a standard oral word frequency reference list (LCMC is based on written texts, and the HSK list is based on pedagogical requirements for learning Chinese in formal contexts, including the need to understand written and oral genres) means that to a certain extent lexical sophistication measures employing reference list methodology in this thesis were comparing apples and oranges. And finally, the exploratory NS data for lexical variety (see 5.3.1.1) was only adequate to show how a researcher might find comparable data; it was not enough data to make solid conclusions about its validity, nor was this technique applied evenly across all CAF methodologies described in the thesis due to time constraints. A more robust study would have sampled NS data in the same context as the NNS data and analysed CAF indices for both sets using identical methods.

**Epistemological issues:** This thesis focused on finding positive examples of language use, rather than analyses of errors in linguistic patterns. It does not tackle the issue of what might be 'missing' from the participant data (e.g. language that is conspicuously absent, patterns of avoidance), or to what degree proficiency can be established if errors are not also taken into account. A particular weakness of the analysis presented here is the lack of accuracy analysis; indeed, the value of a 'CAF' analysis without the 'A' is understandably questionable.

Finally, CAF measures of oral proficiency are not an indication of literacy or cultural proficiency, just one part of comprehensive models of language proficiency; indeed, this thesis focuses even more narrowly on linguistic form, and does not take into account discourse phenomena, functional intention, or concepts that arise from other frameworks in the field of applied linguistics. Thus, researchers, instructors, and learners should realize the limitations of overgeneralizing the results, especially given the current lack of supporting research in the CAL context.

## **5.5 Implications and future work**

### **5.5.1 Importance**

While keeping in mind the relatively small data set and thus exploratory nature of this study, it appears that standards of validity, reliability, and practicality can all be realistically reached with further research into CAF in the CAL context. As more studies come out investigating the specifics of each measurement, validity can be challenged or established; reliability can be assured with gradual agreement on measurement standards and statistical methods; and as more automated tools like Da's vocabulary profiler (introduced in section 4.2.2) are made public and refined, using CAF techniques as part of research and instruction will become more practical.

CAF measures' impact may depend on the field in which they are applied. I believe they are most promising for applied linguistics research (e.g. as research tools to validate pedagogical vocabulary lists, register sty-

listics and their effect on complexity and fluency parameters, and task impact). Better operationalization of parameters discussed in this thesis can have far-ranging impact in different fields; in addition to being an important tool for oral test validation, speech-model operationalization, child language development, researchers are using these concepts to explore translation universals in Chinese texts translated from English (Xiao, 2010) and foreign language order of acquisition studies (Shi, 2002). The impact may be more a dialog than a one-way transferral: for instance, order of acquisition research and other lines of inquiry into accurate lexical use, syntax, and so forth would be a good site for researchers in CAF studies to derive sets of local accuracy indices that are readily comparable to known prescriptive grammar studies. This would effectively provide structure for syntactic variety and accuracy measures in order to speed up operationalization of those aspects in CAF studies.<sup>42</sup>

With respect to oral testing, it is important to study CAF measures to distinguish gains in a more nuanced manner. For instance, studies have shown that differences can be distinguished between the gains of learners in different contexts (study abroad vs. formal classroom learning) but that standardized oral assessments may not adequately represent those differences (Collentine, 2004). Other examples include Iwashita et al.'s (2008) comments about the inevitable reductionism involved in summary scores on tests, and the fact that studies have shown subjective ratings of proficiency depend on different factors at different proficiency levels (e.g. vocabulary and pronunciation contribute more to scores at lower proficiency levels while grammar and fluency contribute more at higher levels). As another example, in terms of the current study, it could be posited that Elliott's discrepancies between high complexity and low fluency might be influenced by his literacy abilities, which he assessed as much higher; perhaps future studies employing CAF methodologies can show that vocabulary knowledge through reading proficiency can influence oral proficiency in certain respects.

---

<sup>42</sup> Cao (2009) provides a detailed overview of Chinese acquisition studies published in Chinese (it does not include studies from the *Journal of Chinese Language Teachers Association*, however), and would be a gold mine for developing CAF indices from established CAL interlanguage research findings.

There are still other, more radical ways, in which CAF studies could change the playing field of oral language testing. As Campbell & Duncan (2007, p.592) put it, “most experience some form of performance anxiety when their own abilities are being evaluated; those who have to evaluate others’ performance routinely find that task especially taxing.” It may be possible to reduce stress for learners and assessors by personalizing tests to learners but simultaneously minimize the efforts assessors need to go through to determine learner proficiency. It may also be possible to get an indication of proficiency when the situation does not favour structured tests such as in-class or standardized assessments. Some of this anxiety, and the potential gaps that might exist in characterizing proficiency via rigid standardized testing, might be addressed via ‘alternative testing’ methodologies such as portfolio development, project work, self-assessment, ‘dynamic assessment’, and ‘can do’ descriptions of language ability (Campbell & Duncan, 2007, p.599). We’ve also seen how participants in the current study, Patsy and Tim, were unable to take standardized tests which assumed literacy skills; their level of proficiency was considerable, but via these traditional methods untestable.

One of the strengths of CAF analysis is that it provides a way to assess unsolicited language; that is, it purports to examine proficiency without tripping over whether or not someone ‘answered the question the way you expected them to.’ Results of CAF studies could be used to inform textbook design based on actual observed order of acquisition or task impact studies, to validate benchmarks in existing or future oral proficiency tests, and provide realistic guidelines for pedagogical word lists. CAF techniques could also be useful in the field of education to determine if a given audio piece or text piece is appropriate for a certain level of student, by examining fluency (speech rate and disfluency, essentially, for NS speech) and complexity (lexical variety and sophistication, syntactic variety and complexity), similar to the grading schemes used to classify books in second or foreign language acquisition contexts.

### 5.5.2 Recommendations and caveats

Currently, the only easily available tools for automation of much of the linguistic segmenting needed to do this work are ICTCLAS and the Stanford Parser. The former appears to do a better job of segmenting, and so should be used as the primary segmenting engine; the latter can then be applied if richer POS and grammar-dependency information is required. For some measurements I had to write specialized scripts (comparing words to reference lists, calculating MATTR), but all source materials are given in the appropriate places in this thesis.

Developing scripts and using the segmenting libraries are not trivial tasks; there are difficulties of character-set encoding (different scripting languages expect different formats, and care must be taken to retain encoding when using editors), text preparation and normalization (so that computer tools act on exactly the right data, ignoring prefatory data, mistaken typing, non-participant data, etc.), transcription standardization (operationalizing definitions of e.g. errors or self-repairs, representing numbers and pauses in consistent ways), and algorithm testing and refactoring (using a simple set of test data to verify that the output matches expectations), to name just a few. These stages of normalization and tool development are often straightforward in isolation but a headache when they all depend on each other for valid analysis. Future work should strive to use clear, standardized parameters and algorithms to establish best practices in this field.

To speed up investigation of individual measures, researchers could consider focusing on just one type of task, and use short excerpts of data taken randomly from multiple test subjects. For involved processes such as developing accuracy and syntactic variety measures one could look at just one data sample in depth; this might help advance ideas in the short term until more ‘plug-and-play’ algorithms are found in the medium-to-long term.

It must be cautioned that oral data sets are tough to work with in any language, but in the CAL context this is especially difficult, as there is little reference data (even the segmenting tools aren’t trained on oral data

sets) and a bias in pedagogy for language conforming to written genre standards, meaning that on available reference scales speech may seem less complex or varied than is posited for a given level. Further, much of the complexity in data transformation arises from the fact that oral data, at least in the present technological milieu, needs to be represented as annotated text to take advantage of high-quality computational linguistics software. While some researchers are beginning to share investigations into obviating the need for manual intervention between oral speech and interpretive software (see Suzuki & Harada, 2004; Bernstein et al., 2010) this work, at least in the field of proficiency assessment, is still very new.

Though many of the techniques explored in the current study are quite involved, and not immediately viable for classroom evaluations, instructors may eventually find some use in these techniques as a way to highlight attention to form for their pupils, by showing fluency numbers from week to week to encourage practice, or showing lexical variety and syntactic complexity in student homework to encourage use of a variety of language and sentence patterns. Swanson & Nolde (2011) explore many options for recording and organizing student speech data for more convenient teacher assessment; if such methods are used frequently in the future, better, more standardized ways of assessing will be even more vital to instructors.

### **5.5.3 Research needed**

As noted in Chapter 2, CAF research is not new in the field of applied linguistics, and there is considerable literature discussing the validity of the various models, how one might structure studies to attempt to test validity, and the general aspects that inform a native speaker's ideas of fluent, proficient language. What is still a nascent enterprise, however, is the application of this robust sub-field in the Chinese-language context.

In addition to the list of gaps in existing research given in section 2.4, the current study highlights the following:

- There is an urgent need for more succinct scholarship in Chinese summarizing the vast amount of thesis and dissertation materials exploring oral proficiency, HSK validation, and interlanguage transcription.
- CAF studies could help answer questions such as: What do people gain by living abroad or studying in various situations? What kinds of output are characteristic of certain learning patterns or contexts? Indeed, we can imagine automated, summarized semantic and grammar analyses of what different test questions or interview prompts elicit.
- The field needs more techniques for examining the correlation between self-assessment, subjective assessment, and objective assessment.
- Cross-language comparisons might be worth investigating as well; how do the numbers generated from CAF measurements differ from existing discussions of oral proficiency or language acquisition in English and other languages? This might point to areas where CAL is unique.
- CAF measures to date have only been able to show quantitative parameters related to form; more research is needed on marrying this with narrative and discourse competence; how well formed are a speaker's thoughts, do the thoughts flow well, what is effect of the style of speech? Yuan (2009) notes that "another popular way to evaluate L2 learners' output is to record the quality and quantity of discourse features in learners' real time interaction" (p. 110). Discourse competency indices do bring up similar issues to any other coding that requires human judgment; e.g. would raters need to be trained in some way to recognize turn-taking and different kinds of clarification or confirmation checks? Coding of discourse features is quite difficult to automate, but you could code them for a representative segment of data, similar to what was suggested for accuracy measures in the above section. This would be an interesting area for future research, especially if it was shown to contribute nuanced information that the traditional CAF measures cannot, or shown to be correlated to traditional CAF measures in some way.

- Multilingual analysis techniques should be developed to retain the richness of the input data; for this thesis, the methodological choice was not to take into account English or code-switching behaviour, ignoring for the sake of convenience the reality that code-switching serves a communicative purpose.
- Once again, studies of NS language to provide baseline data for L2 acquisition research and a platform for comparing NS and NNS data would be invaluable contributions to the field of CAF studies in the CAL context. Similarly, ensuring that L1 baseline data is acquired for each speaker in order to compare with that speaker's L2 data could help tease out what elements of CAF analysis are specific to L2 acquisition vs. specific to a given speaker's general linguistic behaviour.
- Speech-to-text technology should be refined to allow for direct audio-to-annotated text capabilities. This would eliminate the need for much of the technical computer encoding and text processing knowledge that goes into the transcription and transformation stages (not to mention the drudgery).

#### **5.5.4 Extensions**

In preparing this range of CAF tests and observing the results I couldn't help but accumulate ideas about how to extend the work. Naturally, the most important extension might be to carefully record similar results for a larger data set, with the aim of establishing in a more robust way what in this thesis have only been general, tentative statements about the test results.

One helpful extension of this work would be development of a modular test suite, similar to the one put forth on the web by Da (see 4.2.2), to test different parameters. This would require considerable effort to establish e.g. transcription standards for input data. The test suite could include modules for fluency, accuracy, and complexity, and perhaps even eventually pronunciation modules once integrated with speech-to-text capabilities. Ultimately one could imagine a tester watching a video, listening to a student response, or reading a student answer that has already been visually marked up with the computerized assessments; the human tester could then provide a holistic error-check (e.g. "tag 3/8 was marked wrong, but the rest are ok within tolera-



ble limits”). Until that level of accomplishment was achieved, it would still be an invaluable tool for researchers who don’t have time to develop the scripts.

In the long term, we might imagine that CAF studies could lead to a balanced global rating based on a weighted average of various individual measurements (i.e. a mix of complexity, accuracy, and fluency), comparing it with human ratings and self-assessments such as those provided for by the CEFR. For validating theories of language, Housen & Kuiken (2009) echo other researchers in pushing for a wider array of more specific indices to characterize interrelated developments of subsystems. As studies in the CAL context advance they would do well to attend to specific measures and contribute to this wider field, instead of employing opaque or ad-hoc measures such as those described in previous research (see Chapter 2).

## **5.6 Summary**

This chapter moved from covering the observations of the participants in the current study, to examining the study’s results alongside others in the CAL context, and finally to the implications these observations have in the field of applied linguistics. It highlighted the need for more research, and the valuable areas in which such research could bring dividends for those in the fields of applied linguistics and education in particular. It also advanced some lessons learned and possible future work for other researchers interested in pursuing CAF studies in the CAL context.

## Chapter 6. Conclusion

Complexity, accuracy, and fluency (CAF) studies address a number of research needs, and in the Chinese context this is no less true than in other languages. Yuan & Dietrich (2004, p.1) “...call for further experimental and longitudinal studies to verify [their] results by examining learners' performance in real-life situations and the long-term effects of formal instruction in inter-language development;” this study aimed to answer that call, and though it is exploratory, the author hopes it can contribute to the field of proficiency assessment in CAL by stimulating further studies and getting the tools researchers need into their hands more quickly than was previously possible.

This study finds that some complexity and fluency indices are in fact at present operationalizable and unambiguous in the CAL context. The results confirm that of the CAF measures explored in previous research and augmented herein, complexity measures lend themselves best to automated, principled analysis, though there is a long way to go in terms of collecting good reference data to validate the indices. Fluency measures are also very promising for characterizing learner proficiency, especially as technology begins to bridge the gap between audio input and text output. Comparisons of the data with CEFR self-assessment data used as a baseline, and HSK and TOP test results as supplementary reference data, reinforced the case that many of the constructs explored do not just distinguish among learners but also might one day be aligned to standardized tests.

It is important to note that this study was meant to be exploratory and evocative; the amount of data available did not support comprehensive statistical validation of the various CAF measures proposed, nor did the context of the reference data always align perfectly with that of the participant data, necessitating further approximation and the danger of overgeneralization. However, in light of the research goals and of the paucity of studies clearly defining CAF constructs in the CAL context, it is hoped these limitations do not dishonour the efforts in this thesis to discuss greater standardization in CAF studies and provide solid foundations for future researchers to put these constructs to good use in CAL research.

This work is a vital part of a growing field:

- More and more CAL-learner- and native-speaker-produced textual resources are available for research every day. New corpora based on oral data are being created and analyzed, and access is increasingly available to ‘language portfolios’, recorded ‘oral texts’ (including texts like chat conversations, text messages, even perhaps blog posts). This is a rich world of language that begs for the kinds of tools and methodologies explored in this thesis.
- Proficiency assessments are often based on costly or highly specialized testing that may not always be necessary (e.g. for rough estimation in sociolinguistic studies) or possible (e.g. for characterization of language data that was produced at some time in the past, or if the producer is not available to take a test, or if there is no specialized test available for the speaker’s language combination). The methodology in this study addresses this core need for proficiency assessment in standardized but flexible ways.
- Ultimately, this research might provide further insight into what structures exist in instructed and non-instructed learners’ interlanguages, with what frequencies and tendencies, and how that matches with currently understood notions of order of acquisition and pedagogical order of introduction; further CAF research in the Chinese context may even be able to challenge or support fundamental theories of language processing in applied linguistics.

The author earnestly wishes that the best ideas arising from this thesis can be carried forward, and the flaws excised by future research; CAF analysis will doubtless become commonplace in many branches of linguistics and education, and the earlier the field of Chinese applied linguistics adopts coherent standards, the more confidently it can tackle the other questions it faces.

## References

- Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, 25(4), 671-704.
- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17(1), 1-42.
- Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16(4), 449-465.
- BC Ministry of Education (2010). *Final draft: Additional languages - Elementary-secondary curriculum 2010*.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355-377.
- Biq, Y., Tai, J., & Thompson, S. A. (1996). Recent developments in functional approaches to Chinese. In C.-T. J. Huang & A. Y.-H. Li (Eds.), *New horizons in Chinese linguistics* (pp. 97 – 140). Boston: Kluwer Academic.
- Campbell, C., & Duncan, G. (2007). From theory to practice: General trends in foreign language teaching methodology and their influence on language assessment. *Language and Linguistics Compass*, 1(6), 592-611.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.
- Cao, X. (2009). 汉语作为第二语言习得研究中的学习者语言分析方法述评 A review on the methods of analyzing learner language in the studies of the acquisition of CSL. *汉语学习 Chinese Language Learning*, 2009(6).

- Chang, L. (2011). 對應於歐洲共同架構的對外漢語學時建議 (Aligning recommended study time for Chinese as a foreign language to the Common European Framework). Paper presented at: 第一屆東亞華語教學研究生論壇 (*East Asian graduate student forum of teaching Chinese as a second language*). Taipei, Taiwan.
- Chen, F. J. (2009). A multidimensional study of the suffix –men in Chinese: From semantic, discourse, pragmatic and pedagogical perspectives. *Journal of the Chinese Language Teachers Association*, 44(3), 11-42.
- Chen, K.-J., & Liu, S.-H. (1992). Word identification for Mandarin Chinese sentences. *COLING 1992, 14th International Conference on Computational Linguistics* (pp. 102-107). Nantes, France.
- Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, 26(2), 227-248.
- Collentine, J., & Freed, B. F. (2004). Learning context and its effects on second language acquisition. *Studies in Second Language Acquisition*, 26(2), 153-171.
- Council of Europe (2001). *A Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94-100.
- Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics*, 11(2), 183-199.

- Cruickshank, K., & Tsung, L. (2010). Teaching and learning Chinese: A research agenda. In L. Tsung and K. Cruickshank (Eds.), *Learning and teaching Chinese in global contexts: Multimodality and literacy in the new media age* (pp. 213-224). London: Continuum.
- Cui, Y. (2005). 二十年来对外汉语教学研究热点回顾 A brief review of the topics of general interest in the field of teaching Chinese as a Foreign Language in the nearly twenty years. *语言文字应用 Applied Linguistics*, 2005(1), 63-70.
- Da, J. (2005). Reading news for information: How much vocabulary a CFL learner should know. Paper presented at: *International Interdisciplinary Conference on Hànzì rènzhī - How Western learners discover the world of written Chinese*. Gernersheim, Germany.
- Da, Jun (2006). A web-based vocabulary profiler for Chinese language teaching and research. Paper presented at TCLT 4: *Fourth International Conference and Workshops on Technology and Chinese Language Teaching*. Los Angeles, USA.
- Duff, P., Anderson, T., Illynyk, R., Lester (VanGaya), E., Wang, R., Yates, E. (forthcoming). *Learning Chinese: Linguistic, sociocultural, and narrative perspectives*. Mouton de Gruyter.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, U.K: Oxford University Press.
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(03), 299-323.
- Freed, Barbara F. (Ed.) (1995). *Second language acquisition in a study abroad context*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

- Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379-399.
- Guo, X. (2007). 汉语作为第二语言的口语流利性量化测评 (Evaluation of oral fluency in Chinese as a Second Language). *湘潭师范学院学报(社会科学版) Journal of Xiangtan Normal University (Social Science Edition)*, 29(4), 91-94.
- Han, Z. (2003). 韩国学生学习汉语“了”的常见偏误分析 (Analysis of common errors using *le* among Korean students of Chinese). *汉语学习 Chinese Language Learning*, 2003(4), 67-71.
- Hanban (2011). 关于新、旧 HSK 分数对应关系的说明 (Alignment of old and new HSK levels). Retrieved May 15, 2011 from [http://www.hanban.org/news/article/2011-04/14/content\\_248511.htm](http://www.hanban.org/news/article/2011-04/14/content_248511.htm).
- Hanban (n.d.). HSK 项目介绍 (Introduction to the HSK). Retrieved April 25, 2011 from [http://www.hanban.org/tests/node\\_7486.htm](http://www.hanban.org/tests/node_7486.htm).
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in Second Language Acquisition. *Applied Linguistics*, 30(4), 461-473.
- Huang, C.-R., Yo, T.-S., & Šimon, P. (2008). A realistic and robust model for Chinese word segmentation. *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing*.
- Hyland, K., & Zuengler, J. (Eds.). (2009). Complexity, accuracy and fluency (CAF) in second language acquisition research [Special issue]. *Applied Linguistics* 30(4).

- Iwashita, N. (2010). Features of oral proficiency in task performance by EFL and JFL learners. In M. T. Prior (Ed.), *Selected proceedings of the 2008 Second Language Research Forum* (pp. 32-47). Somerville, MA: Cascadia Proceedings Project.
- 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.
- Jiang, W. (2009). Acquisition of word order in Chinese as a foreign language. Berlin: Mouton de Gruyter.
- Jin, H. (2007). Syntactic maturity in second language writings: A case of Chinese as a foreign language (CFL). *Journal of the Chinese Language Teachers Association*, 42(1), 27-54.
- Ke, C. (2005). Patterns of acquisition of Chinese linguistic features by CFL learners. *Journal of the Chinese Language Teachers Association*, 40(1), 1-24.
- Kuiken, F., & Vedder, I. (2008). Task complexity, task characteristics and measures of linguistic performance. 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. 113-125). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.
- Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research. Eurosla Monographs Series, volume 1*. (pp. 81-100). eurosla.org.
- Lazarinis, F., Vilares, J., Tait, J., & Efthimiadis, E. N. (2009). Current research issues and trends in non-English Web searching. *Information Retrieval*, 12(3), 230-250.



- Li, D., & Duff, P. A. (2008). Issues in Chinese Heritage Language education and research at the postsecondary level. In Y. Xiao & A. W. He (Eds.), *Chinese as a heritage language: Fostering rooted world citizenry* (pp. 13-36). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Li, H., Li, H., Fu, N., & He, G. (2007). 汉语常用多义词在中介语语料库中的义项分布及偏误考察 An investigation of the distribution of meaning items and errors of commonly used polysemes found in the Chinese Interlanguage Corpus. *世界漢語教學 Chinese Teaching in the World*, 2007(1), 99-109.
- Li, Y. (2011). Acquisition of the aspectual meanings of negation markers in Mandarin Chinese by English-speaking L2 Chinese learners. *Journal of the Chinese Language Teachers Association*, 46(1), 1-29.
- Liu, S. (2007). *Early vocabulary development in English, Mandarin, and Cantonese: a cross-linguistic study based on CHILDES*. University of Richmond.
- Luo M., Zhang J., Xie O., Huang H., Xie N., & Li Y. (2011). 新汉语水平考试(HSK)质量报告 Report on the quality of new Chinese proficiency test (HSK). *中国考试 China Examinations*, 2011(10), 3-7.
- Magnan, S. S. (1998). Grammar and the ACTFL oral proficiency interview: Discussion and data. *The Modern Language Journal*, 72(3), 266-276.
- Malone, M. E. (2011). Assessment literacy for language educators. *Center for Applied Linguistics Digests*.
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85-104.

- Martínez Baztán, A. (2008). *La evaluación oral: una equivalencia entre las guidelines de ACTFL y algunas escalas del MCER*. (Doctoral dissertation.) Retrieved from Repositorio Institucional de la Universidad de Granada.
- 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.
- Packard, Jerome L. (2000). *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.
- Salamoura, A., & Saville, N. (2010). Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. *Communicative proficiency and linguistic development: intersections between SLA and language testing research. Eurosla Monographs Series, volume 1*. (pp. 101-132).
- SC-TOP (2010). 華語文能力測驗 答客問 (TOCFL Frequently Asked Questions). Retrieved from [http://www.tw.org/top/top\\_intro\\_e.pdf](http://www.tw.org/top/top_intro_e.pdf) on May 4, 2012.
- Shen, H. H. (2005). Linguistic complexity and beginning-level L2 Chinese reading. *Journal of the Chinese Language Teachers Association*, 40(3), 1-28.
- Shi, J. (1998). 外国留学生 22 类现代汉语句式的 习得顺序研究 (Research on the order of acquisition of 22 modern Chinese sentence types by foreign exchange students). *世界汉语教学 Chinese Teaching in the World*, 1998(4).
- Shi, J. (2002). 韩国留学生汉语句式习得的个案研究 (A case study of the acquisition of Chinese sentence patterns by a Korean learner). *世界汉语教学 Chinese Teaching in the World*, 2002(4), 34-42.

- Shneider, G. & Lenz, P. (2003). *European Language Portfolio: Guide for developers*. Strasbourg, Council of Europe. Retrieved June 1, 2012 from [http://www.coe.int/t/DG4/Portfolio/documents\\_intro/Eguide.pdf](http://www.coe.int/t/DG4/Portfolio/documents_intro/Eguide.pdf).
- Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, 17(1), 38-62.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510-532.
- Sun, X. (2008). 准备因素对留学生汉语口语表达的影响 The influence of preparation factor on Chinese oral ability of overseas students. *民族教育研究 Journal of Research on Education for Ethnic Minorities*, 19(4), 89-92.
- Suzuki, M., & Harada, Y. (2004). A common testing framework for measuring spoken language skills of non-native speakers. *IWLLeL 2004: An Interactive Workshop on Language e-Learning* (pp. 115-122).
- Swanson, P. B., & Nolde, P. R. (2011). Assessing student oral language proficiency: Cost-conscious tools, practices & outcomes. *IALLT Journal for Language Learning Technologies*, 41(2), 72-88.
- Taguchi, N. (2008). Building language blocks in L2 Japanese: Chunk learning and the development of complexity and fluency in spoken language production. *Foreign Language Annals*, 41(1), 132-156.
- Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition*, 33(03), 339-372.
- UBC (2006). An introduction to HSK. *Chinese Language Program*. Retrieved July 24, 2011 from [http://www.chinese.arts.ubc.ca/test\\_introduction.htm](http://www.chinese.arts.ubc.ca/test_introduction.htm).

- Vandergrift, L. (2006). Proposal for a common framework for languages for Canada. Retrieved April 25, 2011 from  
[http://www.caslt.org/pdf/Proposal\\_Common%20Framework\\_Reference\\_languages%20for%20Canada\\_PDF\\_Internet\\_e.pdf](http://www.caslt.org/pdf/Proposal_Common%20Framework_Reference_languages%20for%20Canada_PDF_Internet_e.pdf)
- Wang, X. (2011). HSK 语法大纲与教学语法的比较分析 (A comparison of HSK grammar and pedagogical grammar). *语文学刊 Journal of Language and Literature Studies* 2011(6), 1-6.
- Wen, X. (2006). Acquisition sequence of three constructions: An analysis of interlanguage of learners of CFL. *Journal of the Chinese Language Teachers Association*, 43(3), 89-113.
- Wu, S.-L. (2011). Learning to express motion events in an L2: The case of Chinese directional complements. *Language Learning*, 61(2), 414-454.
- Xiao, R. (2010). How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, 15(1), 5-35.
- Xiao, Y. (2004). L2 acquisition of Chinese topic-prominent constructions. *Journal of the Chinese Language Teachers Association*, 39(3), 65-84.
- Xiao, Y. (2011). NP ellipsis in Chinese discourse. *Journal of the Chinese Language Teachers Association*, 46(1), 31-59.
- Yuan, F. (2009). Measuring learner language in L2 Chinese in fluency, accuracy, and complexity. *Journal of the Chinese Language Teachers Association*, 44(3), 109-130.

Yuan, F. (2010). Impacts of task conditions on learner's output in L2 narrative writing. *Journal of the Chinese Language Teachers Association*, 45(1), 67-88.

Yuan, F., & Dietrich, M. S. (2004). Formal instruction, grammatical teachability, and acquisition of CF/SL. *Journal of the Chinese Language Teachers Association*, 39(2), 1-18.

*Zhonghua Advanced New Dictionary [中華高級新詞典]* (2004). Hong Kong: Chunghwa.

## Appendix A – Excerpts from the CEFR for languages: Level descriptors

**Table A-1: CEFR global level descriptions**

Level		Description
Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

**Table A-2: CEFR spoken interaction level descriptions**

Level		Description
Proficient User	C2	I can take part effortlessly in any conversation or discussion and have a good familiarity with idiomatic expressions and colloquialisms. I can express myself fluently and convey finer shades of meaning precisely. If I do have a problem I can backtrack and restructure around the difficulty so smoothly that other people are hardly aware of it.
	C1	I can express myself fluently and spontaneously without much obvious searching for expressions. I can use language flexibly and effectively for social and professional purposes. I can formulate ideas and opinions with precision and relate my contribution skilfully to those of other speakers.
Independent User	B2	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.
	B1	I can deal with most situations likely to arise whilst travelling in an areas where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).
Basic User	A2	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself.
	A1	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics.

**Table A-3: CEFR spoken production level descriptions**

Level		Description
Proficient User	C2	I can present a clear, smoothly-flowing description or argument in a style appropriate to the context and with an effective logical structure which helps the recipient to notice and remember significant points.
	C1	I can present clear, detailed descriptions of complex subjects integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.
Independent User	B2	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.
Basic User	A2	I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.
	A1	I can use simple phrases and sentences to describe where I live and people I know.



## Appendix B – Sample results of grammatical parsers applied to participant interview data

For details on the operation and tag set used by ICTCLAS, please see:

[http://ictclas.org/ictclas\\_files.html](http://ictclas.org/ictclas_files.html)

For details on the operation and tag sets used by the Stanford Parser, please see:

<http://www.cis.upenn.edu/~chinese/>

### B.1 ICTCLAS part-of-speech analysis

#### B.1.1 Input (arbitrary text from Elliott's 2010 interview data)

我...我目前正在学习的是...Berkley 大学所提供的...um...一堂网上...中级...韩国话课。所以，我快要学完，但是学完这个之后，不是可以讲出来的，能...会讲出来的。而是看得懂就好。

#### B.1.2 ICTCLAS output

##### B.1.2.1 Part-of-speech (POS) tagging

我/rr ./wj ./wj ./wj 我/rr 目前/t 在/p 学习/v 的/ude1 是/vshi ./wj ./wj ./wj Berkley/x 大学/n 所/usuo 提供/v 的/ude1 ./wj ./wj ./wj um/x ./wj ./wj ./wj 一/m 堂/q 网上/s ./wj ./wj ./wj 中级/b ./wj ./wj ./wj 韩国/nsf 话/n 课/n 。/wj 所以/c ， /wd 我/rr 快/d 要/v 学/v 完/vi ， /wd 但是/c 学/v 完/vi 这个/rz 之后/f ， /wd 不/d 是 /vshi 可以/v 讲/v 出来/vf 的/ude1 ， /wd 能/v .../ws 会/v 讲/v 出来/vf 的/ude1 。/wj 而是/c 看/v 得/ude3 懂 /v 就/d 好/a 。/wj

## B.2 Stanford Parser grammatical analysis

### B.2.1 Input (arbitrary text from Elliott's 2010 interview data)

我...我目前正在学习的是...Berkley 大学所提供的...um...一堂网上...中级...韩国话课。所以，我快要学完，但是学完这个之后，不是可以讲出来的，能...会讲出来的。而是看得懂就好。

### B.2.2 Stanford Parser output

#### B.2.2.1 POS tagging

我/PN ./AD ./VV ./NN

我/PN 目前/NT 在/AD 学习/VV 的/DEC 是/VC ./NR ./JJ ./NN

Berkley/NR 大学/NN 所/MSP 提供/VV 的/DEC ./NN ./AD ./VV

um/NN ./AD ./VV ./NN

一/CD 堂/NN 网/NN 上/LC ./AD ./VV ./NN

中级/JJ ./NN ./AD ./VV

韩国/NR 话/NN 课/VA 。/PU 所以/AD ， /PU 我/PN 快/VA 要/VV 学完/NN ， /PU 但是/AD 学完/NN 这个/NN

之后/LC ， /PU 不/AD 是/VC 可以/VV 讲/VV 出来/VV 的/DEC ， /PU 能/VV .../PU 会/VV 讲/VV 出来/VV 的

/DEC 。/PU 而/AD 是/VC 看/VV 得/DER 懂/VV 就/AD 好/VA 。/PU

#### B.2.2.2 Grammar tree parsing

```
(ROOT
  (IP
    (NP (PN 我))
    (VP
      (ADVP (AD .))
      (VP (VV .)
        (NP (NN .))))))
```

```
(ROOT
  (IP
    (NP
      (CP
```

```

(IP
  (NP (PN 我))
  (VP
    (NP (NT 目前))
    (ADVP (AD 在))
    (VP (VV 学习))))
  (DEC 的))
(VP (VC 是)
  (NP
    (NP (NR .))
    (ADJP (JJ .))
    (NP (NN .))))))

```

```

(ROOT
  (IP
    (NP
      (CP
        (IP
          (NP
            (NP (NR Berkley))
            (NP (NN 大学))
            (VP (MSP 所)
              (VP (VV 提供))))
          (DEC 的))
        (NP (NN .)))
      (VP
        (ADVP (AD .))
        (VP (VV .))))))

```

```

(ROOT
  (IP
    (NP (NN um))
    (VP
      (ADVP (AD .))
      (VP (VV .)
        (NP (NN .))))))

```

```

(ROOT
  (IP
    (NP
      (QP (CD 一))
      (NP (NN 堂)))
    (VP
      (LCP
        (NP (NN 网))
        (LC 上))
      (ADVP (AD .))
      (VP (VV .)
        (NP (NN .))))))

```

```

(ROOT
  (IP
    (NP
      (ADJP (JJ 中级))

```

```

(NP (NN .)))
(VP
  (ADVP (AD .))
  (VP (VV .))))))
(ROOT
  (IP
    (IP
      (NP
        (NP (NR 韩国))
        (NP (NN 话)))
      (VP (VA 课)))
    (PU 。)
    (IP
      (ADVP (AD 所以))
      (PU , )
      (NP (PN 我))
      (VP
        (VP (VA 快))
        (VP (VV 要)
          (NP (NN 学完))))))
    (PU , )
    (IP
      (ADVP (AD 但是))
      (LCP
        (NP (NN 学完) (NN 这个))
        (LC 之后))
      (PU , )
      (VP
        (ADVP (AD 不))
        (VP (VC 是)
          (NP
            (CP
              (IP
                (VP (VV 可以)
                  (VP
                    (VP (VV 讲)
                      (NP
                        (CP
                          (IP
                            (VP (VV 出来)))
                            (DEC 的))))))
                (PU , )
                (VP (VV 能)
                  (VP
                    (LST (PU ...))
                    (VP (VV 会)
                      (VP
                        (VCD (VV 讲) (VV 出来))))))))))
                (DEC 的))))))
      (PU 。)
      (IP
        (ADVP (AD 而))
        (VP (VC 是)

```

```

(IP
  (VP (VV 看) (DER 得)
    (VP (VV 懂)
      (VP
        (ADVP (AD 就))
        (VP (VA 好))))))
(PU 。)))

```

### B.2.2.3 Typed dependencies, collapsed

```
nsubj(. -3, 我-1)
```

```

nsubj(学习-4, 我-1)
dep(学习-4, 目前-2)
advmod(学习-4, 在-3)
dep(的-5, 学习-4)
top(是-6, 的-5)
root(ROOT-0, 是-6)

```

```

nn(大学-2, Berkley-1)
nsubj(提供-4, 大学-2)
prtmod(提供-4, 所-3)
dep(的-5, 提供-4)
rcmod(. -6, 的-5)

```

```
nsubj(. -3, um-1)
```

```

nummod(堂-2, 一-1)
nsubj(. -6, 堂-2)
lobj(上-4, 网-3)
loc(. -6, 上-4)

```

```
amod(. -2, 中级-1)
```

```

nn(话-2, 韩国-1)
nsubj(课-3, 话-2)
root(ROOT-0, 课-3)
advmod(快-8, 所以-5)
nsubj(快-8, 我-7)
dep(课-3, 快-8)
conj(要-9, 快-8)
dep(快-8, 要-9)
dobj(要-9, 学完-10)
advmod(是-18, 但是-12)
nn(这个-14, 学完-13)
lobj(之后-15, 这个-14)
loc(是-18, 之后-15)

```

neg(是-18, 不-17)  
 dep(课-3, 是-18)  
 mmod(讲-20, 可以-19)  
 dep(的-29, 讲-20)  
 dep(的-22, 出来-21)  
 dobj(讲-20, 的-22)  
 mmod(讲-27, 能-24)  
 mmod(讲-27, 会-26)  
 conj(讲-20, 讲-27)  
 comod(讲-27, 出来-28)  
 attr(是-18, 的-29)  
 advmod(是-32, 而-31)  
 dep(课-3, 是-32)  
 dep(好-37, 看-33)  
 dep(好-37, 得-34)  
 mmod(好-37, 懂-35)  
 advmod(好-37, 就-36)  
 ccomp(是-32, 好-37)

### B.2.3 Input (same text from Elliott's 2010 interview data, cleaned version)

我目前在学习的是 Berkley 大学所提供的一堂网上中级韩国话课。所以，我快要学完，但是学完这个之后，不是可以讲出来的，能，会讲出来的。而是看得懂就好。

### B.2.4 Stanford Parser output

#### B.2.4.1 POS tagging

我/PN 目前/NT 在/AD 学习/VV 的/DEC 是/VC Berkley/NR 大学/NN 所/MSP 提供/VV 的/DEC 一/CD 堂/NN 网  
 /NN 上/LC 中级/JJ 韩国/NR 话/NN 课/NN 。/PU 所以/AD ， /PU 我/PN 快/VA 要/VV 学完/NN ， /PU 但是/AD  
 学完/NN 这个/NN 之后/LC ， /PU 不/AD 是/VC 可以/VV 讲/VV 出来/VV 的/DEC ， /PU 能/VV ， /PU 会/VV 讲  
 /VV 出来/VV 的/DEC 。/PU 而/AD 是/VC 看/VV 得/DER 懂/VV 就/AD 好/VA 。/PU

#### B.2.4.2 Grammar tree parsing

```

(ROOT
  (IP
    (IP

```

(NP  
   (CP  
     (IP  
       (NP (PN 我))  
       (VP  
         (NP (NT 目前))  
         (ADVP (AD 在))  
         (VP (VV 学习)))  
       (DEC 的)))  
     (VP (VC 是))  
     (NP  
       (LCP  
         (NP  
           (CP  
             (IP  
               (NP  
                 (NP (NR Berkley))  
                 (NP (NN 大学)))  
               (VP (MSP 所)  
                 (VP (VV 提供))))  
             (DEC 的))  
             (QP (CD 一))  
             (NP (NN 堂) (NN 网)))  
           (LC 上))  
           (ADJP (JJ 中级))  
           (NP (NR 韩国) (NN 话) (NN 课))))))  
   (PU 。)  
   (IP  
     (ADVP (AD 所以))  
     (PU , )  
     (NP (PN 我))  
     (VP  
       (VP (VA 快))  
       (VP (VV 要)  
         (NP (NN 学完))))))  
   (PU , )  
   (IP  
     (ADVP (AD 但是))  
     (LCP  
       (NP (NN 学完) (NN 这个))  
       (LC 之后))  
     (PU , )  
     (VP  
       (ADVP (AD 不))  
       (VP (VC 是))  
       (NP  
         (CP  
           (IP  
             (VP (VV 可以))  
             (VP  
               (VP (VV 讲)  
                 (NP  
                   (CP  
                     (IP

```

                (VP (VV 出来)))
            (DEC 的))))
        (PU , )
        (VP (VV 能))
        (PU , )
        (VP (VV 会)
            (VP
                (VCD (VV 讲) (VV 出来))))))
    (DEC 的))))))
(PU 。)
(IP
    (ADVP (AD 而))
    (VP (VC 是)
        (IP
            (VP (VV 看) (DER 得)
                (VP (VV 懂)
                    (VP
                        (ADVP (AD 就))
                        (VP (VA 好)))))))
        (PU 。)))

```

### B.2.4.3 Typed dependencies, collapsed

```

nsubj(学习-4, 我-1)
dep(学习-4, 目前-2)
advmod(学习-4, 在-3)
dep(的-5, 学习-4)
top(是-6, 的-5)
root(ROOT-0, 是-6)
nn(大学-8, Berkley-7)
nsubj(提供-10, 大学-8)
prtmod(提供-10, 所-9)
dep(的-11, 提供-10)
rcmod(网-14, 的-11)
rcmod(课-19, 的-11)
nummod(网-14, 一-12)
nn(网-14, 堂-13)
lobj(上-15, 网-14)
dep(课-19, 上-15)
amod(课-19, 中级-16)
nn(课-19, 韩国-17)
nn(课-19, 话-18)
attr(是-6, 课-19)
advmod(快-24, 所以-21)
nsubj(快-24, 我-23)
dep(是-6, 快-24)
conj(要-25, 快-24)
dep(快-24, 要-25)
dobj(要-25, 学完-26)
advmod(是-34, 但是-28)

```



nn(这个-30, 学完-29)  
lobj(之后-31, 这个-30)  
loc(是-34, 之后-31)  
neg(是-34, 不-33)  
dep(是-6, 是-34)  
mmod(讲-36, 可以-35)  
dep(的-45, 讲-36)  
dep(的-38, 出来-37)  
dobj(讲-36, 的-38)  
conj(讲-36, 能-40)  
mmod(讲-43, 会-42)  
conj(讲-36, 讲-43)  
comod(讲-43, 出来-44)  
attr(是-34, 的-45)  
advmod(是-48, 而-47)  
dep(是-6, 是-48)  
dep(好-53, 看-49)  
dep(好-53, 得-50)  
mmod(好-53, 懂-51)  
advmod(好-53, 就-52)  
ccomp(是-48, 好-53)