Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research

Xiaobin Chen

Advisors
Prof. Dr. Detmar Meurers
Prof. Dr. Katharina Scheiter
A dissertation presented for the degree of
Doctor of Philosophy

LEAD Graduate School & Research Network
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen
Germany
December 27, 2018

Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research

Dissertation
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Xiaobin Chen

aus

Guangdong, China

2018

Gedrukt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan(in): Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter(in): Prof. Dr. Detmar Meurers
Mitberichterstatter(in): Prof. Dr. Katharina Scheiter

Tag der mündlichen Prüfung: 14.12.2018

Tübingen

To my wife Chang Wei,
and our children
Danning and Yuanning.

Declarations

Erklärung bzgl. anderer Promotionsverfahren
Hiermit erkläre ich, dass ich mich keinen anderen Promotionsverfahren oder entsprechenden Prüfungsverfahren unterzogen habe.

Erkliärung bzgl. Verwendung von Hilfsmitteln
Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel: „Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research“ selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Erklärung bzgl. kommerzieller Unterstützung
Hiermit erkläre ich, dass die Gelegenheit zum vorliegenden Promotionsverfahren nicht kommerziell vermittelt wurde. Ich habe die Dissertation ohne kommerzielle Hilfe von Dritten verfasst und keine Organisation eingeschalten, die gegen Entgelt Betreuer für die Anfertigung von Dissertationen gesucht hat oder die für mich die mir obliegenden Prüfungsleistungen ganz oder teilweise erledigt hat.
Hiermit bestätige ich, dass es mir die Rechtsfolge der Inanspruchnahme eines gewerblichen Promotionsvermittlers und die Rechtsfolge bei Unwahrhaftigkeiten in dieser Erklärung (Ausschluss der Annahme als Doktorand, Ausschluss der Zulassung zum Promotionsverfahren, Abbruch des Promotionsverfahrens und Rücknahme des erlangten Grades wegen Täuschung gemäß § 21) bekannt ist.

Erklärung bzgl. Verurteilungen
Hiermit erkläre ich wahrheitsgemäß und vollständing, dass keine wissenschaftsbezogene strafrechtliche Verurteilungen, Disziplinarmaßnahmen und anhängige Disziplinarverfahren gegen mich laufen.

Xiaobin Chen
December 27, 2018, Tübingen

Acknowledgments

This dissertation would not have been possible without the considerable support and careful supervision from my supervisor Prof. Detmar Meurers, who granted me maximum academic freedom to pursue the interesting topics in this research and provided me with numerous invaluable advice on various issues along the way. His great passion and serious attitude towards scientific inquiry have been and will always be an aspiration of my research career. A million thanks to Detmar.

I owe my deepest gratitude to my family who have always been supportive to my work. I would like to thank my parents, Chen Huihua and Cai Huihua, for bearing with the deprivation of the happiness they had been enjoying living with their favorite grand-daughter before I decided to move far away to Germany to do my PhD with the family.

I am heartily thankful to my wife Chang Wei who gave up a decent job in a Chinese university to become a full-time housewife so as to support my research. Because of her great sacrifice and goodwill, I am able to come to the warmth of a well managed home with two happy children and a delicious meal everyday after a hard day's work. It is really the greatest comfort in the whole world.

My gratitude also goes to my parents-in-law Chang Jingmin and Zhang Yanqing for every support they provided us with. Life would have been harder without their generosity and trust.

I am also indebted to many of my advisors, colleagues, and friends, in particular Patrick Rebuschat, Katharina Scheiter, Michael Grosz, Björn Rudzewitz, Maria Chinkina, Simón Ruiz, Sabrina Galasso, and Zarah Weiß among a lot others. I am forever grateful to the interesting discussions with them on issues related or unrelated to research at the lunch breaks and on the other occasions. The insights I gained from these wonderful colleagues are what makes me develop into a competent researcher and critical thinker.

Last but by no means the least, I would like to thank the LEAD Graduate School & Research Network of the University of Tübingen, who funded my PhD and provided me with a wonderful interdisciplinary research environment, as well as gener-

ous fundings for conferences and intramural projects to make my research possible. My special thanks go to Mareike Bierlich, Sophie Freitag, Katharina Lichtenberger, and the other members of the Scientific Coordination team and the steering board of LEAD for their support and management of the projects leading to this dissertation.

Funding

This research was funded by the LEAD Graduate School & Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments.

Abstract

The construct of complexity, together with accuracy and fluency have become the central foci of language learning research in recent years. This dissertation focuses on complexity, a multidimensional construct that has its own working mechanism, cognitive and psycholinguistic processes, and developmental dynamics. Six studies revolving around complexity, including its conceptualization, automatic measurement, and application in language acquisition research are reported.

The basis of these studies is the automatic multidimensional analysis of linguistic complexity, which was implemented into a Web platform called Common Text Analysis Platform by making use of state-of-the-art Natural Language Processing (NLP) technologies . The system provides a rich set of complexity measures that are easily accessible by normal users and supports collaborative development of complexity feature extractors.

An application study zooming into characterizing the text-level readability with the word-level feature of lexical frequency is reported next. It was found that the lexical complexity measure of word frequency was highly predictive of text readability. Another application study focuses on investigating the developmental interrelationship between complexity and accuracy, an issue that conflicting theories and research results have been reported. Our findings support the simultaneous development account.

The other few studies are about applying automatic complexity analysis to promote language development, which involves analyzing both learning input and learner production, as well as linking the two spaces. We first proposed and validated the approach to link input and production with complexity feature vector distances. Then the ICALL system SyB implementing the approach was developed and demonstrated. An effective test of the system was conducted with a randomized control experiment that tested the effects of different levels of input challenge on L2 development. Results of the experiment supported the comprehensible input hypothesis in Second Language Acquisition (SLA) and provided an automatizable operationalization of the theory.

The series of studies in this dissertation demonstrates how language learning research can benefit from NLP technologies. On the other hand, it also demonstrates how these technologies can be applied to build practical language learning systems based on solid theoretical and research foundations in SLA.

Publications

Part of the contents of this dissertation is based on the following publications and submitted manuscripts.

- Chen, X., Meurers, D., and Rebuschat, P. (Submitted-a). Investigating Krashen's *i + 1*: An experimental ICALL study on the development of L2 complexity.

- Chen, X., Weiß, Z., and Meurers, D. (Submitted-b). Is there a developmental trade-off between complexity and accuracy in L1 and L2 acquisition?.

- Chen, X. and Meurers, D. (2018a). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, In press.

- Chen, X. and Meurers, D. (2018b). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

- Chen, X. and Meurers, D. (2017a). Challenging learners in their individual Zone of Proximal Development using pedagogic developmental benchmarks of syntactic complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa*, pages 8–17, Gothenburg, Sweden, 22nd May. Linköping University Electronic Press, Linköpingsuniversitet.

- Chen, X. and Meurers, D. (2016b). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity at COLING*, pages 113–119, Osaka, Japan, 11th December. The International Committee on Computational Linguistics.

- Chen, X. and Meurers, D. (2016a). Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP*

for Building Educational Applications at NAACL, pages 84–94, San Diego, CA. Association for Computational Linguistics.

Contents

List of Figures

List of Tables

Acronyms

AAE Aggregate Analysis Engine

AE Analysis Engine

BNC British National Corpus

CAF Complexity, Accuracy, and Fluency

CALL Computer Assisted Language Learning

CD Contextual Diversity

CEFR Common European Framework of Reference for Languages

CFA Confirmatory Factor Analysis

CH Cognition Hypothesis

CIPW Complex Input Primed Writing

CTAP Common Text Analysis Platform

CV Cross Validation

CW Continuation Writing

EFL English as a Foreign Language

ESL English as a Second Language

ETS Educational Testing Service

FLAIR Form-Focused Linguistically Aware Information Retrieval

GUI Graphical User Interface

GWT Google Web Toolkit

ICALL Intelligent Computer Assisted Language Learning

IH Input Hypothesis

KNN K-Nearest Neighbors

L1 First Language

L2 Second Language

LCA Lexical Complexity Analyzer

LDT Lexical Decision Tasks

ML Machine Learning

MLTU Mean Length of T-unit

MTLD Measure of Textual Lexical Diversity

NGSL New General Service List

NLP Natural Language Processing

POS Part-Of-Speech

SCA Syntactic Complexity Analyzer

SD Standard Deviation

SEM Structural Equation Modeling

SLA Second Language Acquisition

SyB Syntactic Benchmark

TAALES Tool for the Automatic Analysis of LExical Sophistication

TACIT Text Analysis, Crawling, and Interpretation Tool

TBLT Task-Based Language Teaching

TL Target Language

TOH Trade-Off Hypothesis

TTR Type-Token Ratio

UIMA Unstructured Information MAnagement

ZPD Zone of Proximal Development

Chapter 1

Introduction

Originating from First Language (L1) acquisition research in the 1970s (Brown, 1973; Hunt, 1965), the construct of complexity, together with the notions of accuracy and fluency have become major research variables in applied linguistics (Housen et al., 2009). The three constructs are collectively known as Complexity, Accuracy, and Fluency (CAF) and have been systematically used to study the effects of learning factors, such as task conditions, individual differences, and types of instructions on the performance and/or attainment of the Second Language (L2) (e.g. Skehan, 1989, 1998, 2009; Norris and Ortega, 2009; Spada and Tomita, 2010; Ferraris, 2012). In these studies, CAF usually feature as dependent variables whose variation is attributable to elements that would affect the learners' language production. Yet another strand of research looks at CAF as independent variables and uses them to evaluate the learners' L2 proficiency or model its longitudinal development (e.g. Ortega, 2003; Byrnes, 2009; De Clercq and Housen, 2017; Polat and Kim, 2014; Vyatkina et al., 2015). CAF, in these cases, have become the primary foci of SLA research (Housen et al., 2012, 2009).

Any attempt to give general definitions to CAF would run the risk of oversimplification because all the CAF constructs may contain multiple dimensions and can be approached from multiple perspectives, depending on the research questions and analytical needs. However, there are also some common perspectives and definitions that researchers generally agree upon. For example, complexity, or linguistic complexity to be more specific, can be generally defined as the variedness, elaborateness, and inter-relatedness of the linguistic components in language production (Ellis, 2003; Wolfe-Quintero et al., 1998). Accuracy refers to the non-nativelike production error rate (Polio and Shea, 2014) and fluency to nativelike production speed and smoothness (Lennon, 1990; Kormos and Dénes, 2004; Sato, 2014). CAF as defined in these ways have been found to be able to account for L2 performance as dis-

tinct but interrelated constructs (Norris and Ortega, 2009; Skehan and Foster, 1997; Robinson, 2001). Theoretically, they are also claimed to be related to L2 proficiency and development (Towell, 2012; Housen et al., 2012).

Among the CAF triad, complexity is undoubtedly the most controversial, most complicated and most researched construct (Pallotti, 2009; Bulté and Housen, 2012; Housen et al., 2009). It is 'a multifaceted, multidimensional, and multilayered phenomenon that has cognitive, pedagogical, and linguistic dimensions, developmental and performance aspects, and can manifest itself at all levels of language structure, learning, and use.' (Housen, 2014, p. 63). A plethora of research has been devoted to the conceptualization, measurement, and application of complexity for L2 learning research. This research ranges from investigating the effects of learning tasks or genre on the complexity of the learner's L2 production (e.g. Robinson, 2011; Michel et al., 2007; Tabari, 2016; Foster and Skehan, 1996; Yuan and Ellis, 2003; Yang et al., 2015; Yoon and Polio, 2017; Alexopoulou et al., 2017), to the longitudinal development of L2 linguistic complexity (e.g. Bulté and Housen, 2014; Vyatkina, 2012; Bulté et al., 2008; Vyatkina et al., 2015; Larsen-Freeman, 2006; Yoon and Polio, 2017; Ortega, 2015) and its relationship with the other two constructs of the CAF triad (Spoelman and Verspoor, 2010; Vercellotti, 2017; Robinson, 2005, 2001). Practically, complexity has also been used to analyze both learning input and learner production for purposes such as assessing text readability (Benjamin, 2012; Collins-Thompson, 2014; Vajjala and Meurers, 2012), evaluating the interaction between target structure teachability and teaching methodologies (DeKeyser, 1998; Housen et al., 2005; Spada and Tomita, 2010), or assessment of essay quality (Yang et al., 2015; Taguchi et al., 2013; Crossley et al., 2016), etc. The pervasive use of the complexity construct in previous research has proved its usefulness as an important instrument of SLA research. It has helped theorists to discover the process and working mechanism of SLA and language teaching practitioners to implement more efficient L2 teaching methodologies.

Fruitful results in this line of research notwithstanding, theoretical and practical questions on complexity still baffle researchers. For example, although the multidimensionality of the construct has been well documented in a number of publications (Bulté and Housen, 2012; Housen et al., 2012; Wolfe-Quintero et al., 1998), the majority of studies still adopted a reductionistic view on complexity. Practically, despite the fact that complexity can be operationally measured at all levels of linguistic representations, including lexical, morphological, phonological, syntactic, and discoursal levels, resulting in a large number of indexes for its measurement, most studies to date still use a few measures to represent the construct, probably

'due to the lack of adequate computational tools for automatic complexity measurement and the labor-intensiveness of manual computation' (Bulté and Housen, 2012, p. 34). Besides, in terms of L2 teaching application, it is still difficult to implement the current research findings because of the observational nature of this research. The bulk of the previous research in this area either analyzed the complexity of learning input or learner production separately, or observed the development of L2 complexity in naturally occurring cross-sectional or longitudinal data. Few studies have ever tackled the link between the complexity of learning input and learner production, least the effects of complex input on L2 development. As a result, it is difficult for L2 instructors, teaching material and system designers to apply the research results to actual teaching practice or the development of instructional material and computational system to help the learners' promote their L2 proficiency.

This dissertation tries to address some of the research gaps identifiable from previous research by presenting projects revolving around the topic of linguistic complexity. Specifically, these projects include work on the automatic analysis of linguistic complexity (Chapter 2), the application of complexity analysis to analyze learning input for readability assessment of reading texts (Chapter 3), to analyze learner output for researching language development (Chapter 4), and to simultaneously analyze both learning input and learner production for examining the effects of complex input on L2 proficiency development (Chapters 5–7). In what follows, we will first review how complexity has been approached by previous studies and clarify the definition of the construct in this dissertation. Then we will discuss how complexity can be measured, especially automatically with state-of-the-art NLP tools. Review of the application of linguistic complexity in SLA research will follow and the research gaps identified. The chapter ends with an overview of the studies included in the dissertation.

1.1 The multidimensionality of complexity

The multidimensionality of L2 complexity has been well-documented in a number of publications (e.g. Housen, 2014; Housen et al., 2012; Bulté and Housen, 2012; Pallotti, 2009, 2015). Bulté and Housen (2012) provides a comprehensive taxonomy of L2 complexity, starting from differentiating between the two general constructs of *absolute* and *relative* complexity and going down to the more fine-grained levels of linguistic manifestation (Figure 1.1). The authors further suggested that complexity analysis be approached from three different levels: (a) an abstract theoretical level of cognitive constructs, (b) a more concrete observational level of behavioral constructs,

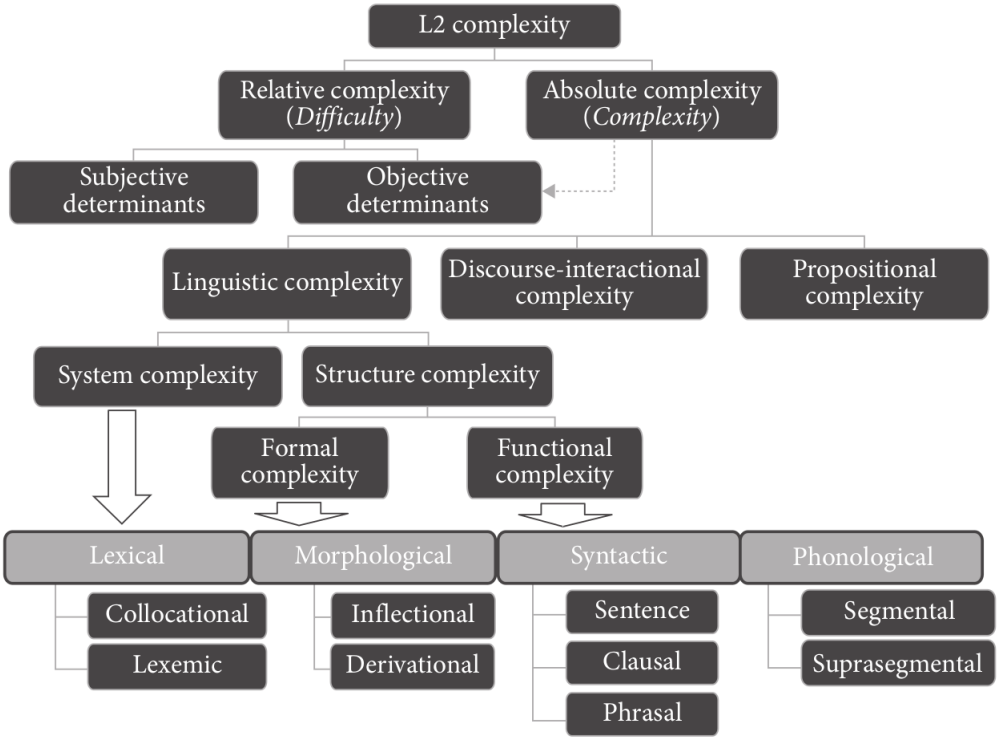

Figure 1.1: A taxonomy of complexity constructs, reproduced from Bulté and Housen (2012, p. 23)

and (c) an operational level of statistical constructs (pp. 26–27). Based on Bulté and Housen's (2012) taxonomy and analytical framework, the following paragraphs will try to disentangle the multidimensionality of the complexity construct.

Two senses for the entry 'complex' are given in the Merriam-Webster dictionary: (a) composed of two or more parts, (b) hard to separate, analyze, or solve. The former sense is related to the compositional makeup of an entity or phenomenon, while the latter is more about the perceived difficulty to understand something. This points to the necessity to distinguish between objective and subjective complexity, or in the more generally agreed terms *absolute* and *relative* complexity in language learning (Miestami, 2008; Bulté and Housen, 2012; Kusters, 2008). Relative complexity is also referred to as *cognitive* complexity or simply *difficulty* because it has to do with the mental ease or difficulty to process, learn or use certain linguistic structures (Hulstijn and de Graff, 1994). Absolute complexity, or linguistic complexity or even simply complexity, on the other hand, mainly concerns about the characteristics or the linguistic features of the language production. Following Rescher's (1998) philosophical view on complexity in general, Ellis (2003) defined linguistic complexity as the *variedness* and *elaborateness* of language production. As a result, suffice it to say that the subcomponents of linguistic complexity con-

sist of the number of distinct linguistic structures at multiple linguistic levels and the number of interrelated connections of these structures to each other (Bulté and Housen, 2012).

Even if we focus on the part of the complexity construct that does not take into account the interaction between the language user/learner and the language production, absolute complexity in a broader sense is still multicomponential. It consists of (a) *propositional complexity*, which refers to the number of information units conveyed in a production segment (Ellis and Barkhuizen, 2005), (b) *discourse-interactional complexity*, which is characterized by interactional moves and their nature in conversational discourse (Pallotti, 2008), and (c) *linguistic complexity*, which is determined by the breadth of the linguistic system (what is available) and the depth of linguistic structures (in how many situations a feature can be used) (Bulté and Housen, 2012). With respect to L2 acquisition, system and structure complexity can both be used to analyze not only the complexity of learning input, but also that of the learner production. Linguistically and on an operational level, complexity can be manifested on all levels of linguistic representation, including phonological, lexical, morphosyntactic, and discoursal levels.

Besides distinguishing L2 complexity on the cognitive and linguistic dimensions, another dimension of complexity that interests SLA researchers and L2 teaching practitioners is its development. Besides being used to gauge proficiency and describe performance—the more static or snapshot views of the construct, complexity is also often used to benchmark L2 development (Ortega, 2012). Pallotti (2015) identifies 'developmental complexity' as one of the three main meanings of complexity, with the other two being structural and cognitive complexity similar to what Bulté and Housen (2012) proposed. It is understood that L2 development is not part of the complexity construct, but complexity measures can empirically be related to developmental stages. Complex L2 production is often associated with later developmental stages or higher L2 proficiency because of the expansion of the learners' linguistic repertoire and capability to use a wider range of linguistic resources (Ortega, 2015). For example, Foster and Skehan (1996) observed progressive elaboration and variation of the learners' production and attributed the phenomena to the development of their interlanguage. The developmental view on L2 complexity is important for studying SLA.

In this dissertation, complexity is approached from all three major dimensions: structural, cognitive, and developmental. An automatic complexity analysis system (Chapter 2) was developed to extract structural complexity features from texts. Lexical complexity was used to predict the cognitive demand for understanding read-

ing texts (Chapter 3). Developmentally, the effects of complex input on L2 performance was tested (Chapters 5 and 7) and the findings implemented into a prototype of an Intelligent Computer Assisted Language Learning (ICALL) system (Chapter 6). An intervention study that tackled both the cognitive and developmental dimensions was also conducted (Chapter 7).

1.2 Measuring linguistic complexity

It is obvious that complexity measurement happens on the operational level and it can be done at all levels of linguistic representation (see Figure 1.1). Generally, complexity is measured by counting the occurrences of linguistic items, calculating their variation rate, relative density, or the percentage of items that are considered complex (e.g. less frequent or acquired later) in reference sources. Skehan (2009) distinguished between *text-internal* and *text-external* complexity measures. The former refers to measures that can be calculated by looking at the text itself, while the latter requires some external references like word frequency lists for the calculation. Text-internal measures are referred to by different researchers as *richness*, *diversity*, or *variation* measures and text-external ones as *sophistication* or *rareness* measures. These measures can be extracted from the morphological, lexical, syntactic, and discoursal levels, resulting in indexes of structural counts, diversity, density, and sophistication.

Structural count measures refer to the tallying of the number of times certain linguistic elements occur in the text/speech being analyzed. For instance, in terms of lexis, the number of word tokens of different Part-Of-Speech (POS) can be counted. Other related measures include indexes like mean length of words in syllables/characters or number of sentences and their average length, etc. Diversity measures count the number of unique tokens or the ratio of unique types to word tokens, or Type-Token Ratio (TTR). These measures are sensitive to text length, but the effect can be accounted for with normalized variants of the TTR such as the G index (Guiraud, 1954), vocd-D (Malvern and Richards, 2002; Malvern et al., 2004), mean segmental TTR, the Uber index (Dugast, 1979), or the MTLD (McCarthy and Jarvis, 2010) etc. Density measures are calculation of how often certain constructs occur in relation to the other constructs. For example, lexical (as opposed to grammatical) density refers to the percentage of words that are lexical words, i.e. nouns, adjectives, verbs, and adverbs with an adjectival base (Lu, 2012). Examples of other density measures include ratio of verb type count to lexical token count (used in Casanave, 1994; Engber, 1995; Linnarud, 1986; Hyltenstam, 1988),

ratio of modifier count (adjectives and adverbs) to lexical token count (first used in McClure, 1991), and on the syntactic level, subordinate clause ratio, passive sentence ratio and so on (Wolfe-Quintero et al., 1998). Sophistication measures refer to the calculation of sophisticated linguistic components in the text with the assistance of external references such as word frequency lists of normed language use. Measures of this type used in previous research include number of sophisticated words per 100 words, mean frequency of words in frequency lists, percentage of 'difficult' words and so on. Sophisticated or difficult words are defined as words that are less frequent, acquired late in life by native speakers, or appear late in teaching materials.

Combining the linguistic levels with these measurement methods, a large number of measures can be reached. For example, Wolfe-Quintero et al. (1998) collected and reviewed more than 100 measures from 39 studies of L2 development. Bulté and Housen (2012) includes a list of 40 grammatical and lexical complexity measures (p. 30). Both Housen (2015) and Vajjala (2015) reported over 200 indexes for measuring L2 complexity and doing readability assessment for language education purposes. Studies reported in this dissertation (Chapters 4, 5 and 7) used over 570 English complexity measures (see Appendix A for a full list). It is very difficult, if not at all impossible, to manually extract all these measures from corpora of non-trivial sizes. Bulté and Housen (2012) thus reported that most previous studies on linguistic complexity used no more than three measures, although there was no shortage of complexity measures in SLA studies. The problem with using a few measures to represent the complexity construct is that by leaving out most of the other aspects, the representation thus created is either biased or incomplete. If a study could not find an effect or causes of an effect, it would not be easy to conclude on the findings because such conclusions are prone to the argument-from-ignorance fallacy.

However, the once-common practice of using a few representative or 'best' measures for different research needs is changing with the development of computational tools for complexity analysis. Systems making use of the latest NLP technologies have been developed for extracting various complexity measures from texts. For instance, Lu's (2012) Lexical Complexity Analyzer (LCA) is capable of calculating 25 metrics on the 3 dimensions of lexical richness, density, and sophistication. He also developed an L2 Syntactic Complexity Analyzer (SCA) (Lu, 2010) that aims at automating the analysis of syntactic complexity of written English production by advanced learners. It is capable of extracting 14 syntactic complexity measures. Kyle and Crossley (2015) also created the Tool for the Automatic Analysis of LExical Sophistication (TAALES) that calculates 135 lexical complexity indices, most of

which with reference to word frequency or psycholinguistic word information from external sources. CohMetrix (McNamara et al., 2014) is another tool for complexity feature extraction that mainly focuses on the discoursal measures of cohesion, including indices such as global/local content word overlap, argument (nouns and pronouns) overlap, syntactic similarity by means of counting common nodes in parse trees and so on. The Common Text Analysis Platform (CTAP) (Chen and Meurers, 2016b, see Chapter 2) developed by the author extends the functionalities of previous complexity analysis systems by adding modules for corpus management, feature selection, and results visualization. The system was implemented as a Web application with a friendly user interface that makes it easy to be used by linguists and researchers who are not familiar with computer programming or NLP technologies. The first release of the system provides over 170 lexical and syntactic measures. The integration of the full set of over 570 complexity measures listed in Appendix A is underway.

Technologically speaking, the computational tools required for extracting complexity measures from texts are generally available from the field of NLP. For lexical complexity measures, common NLP tools such as sentence segmenters, tokenizers, lemmatizers, and POS taggers are usually sufficient for measures of lexical component counts, lexical variation, lexical density, and lexical sophistication. Syntactic complexity measures would require tools such as parsers and tree pattern matchers. Parsers are tools for creating structural or relational representation of the components or entities of a sentence, the results of which are usually represented as trees, a data structure that is processable by the computer. Tree pattern matchers can then be used to identify the syntactic components or structures that are of interest to the researcher. For example, patterns can be used to extract subtrees like subordinate clauses, T-units, and complex nominals, etc.

The automatic analysis of linguistic complexity with NLP is applicable to the analysis of both authentic and learner-produced texts. Since modern NLP tools have reached very high level of accuracy, the analysis of well-formed authentic texts does not pose too much of a challenge to automatic complexity analysis. However, when it comes to learner produced texts, things can get complicated. Because the current NLP tools usually use probabilistic models trained with corpus of well-formed language, they might not work equally well with learner output that contains a lot of grammatical errors and/or misspelled words. A number of studies have documented the challenges parsers tend to meet when it comes to parsing learner language (Geertzen et al., 2013; Ott and Ziai, 2010; Krivanek and Meurers, 2011). For example, Geertzen et al. (2013) found that the Standford CoreNLP parser (Man-

ning et al., 2014) worked well for morphological errors, but struggled with more complex errors. Results from Ott and Ziai (2010) showed that when processing L2 German data, a canonical dependency parser became less reliable if key elements of a sentence (e.g. the main verb) are missing. Although Lu (2012) confirmed the high reliability of his tool for automatic syntactic complexity analysis with writings by English-major students from nine different Chinese universities, these students were considered upper-intermediate learners of English so the results should not be easily generalized to learners of lower proficiency levels. Consequently, depending on the analysis needs and the proficiency of the students who produced the texts, it is suggested that the NLP tools be adapted to learner data, for example, by using an annotation scheme that is sensible to grammar mistakes commonly found in learner writings (Cahill, 2015).

Besides Appendix A of the current thesis, inventories of complexity measures used in previous research can be found in Wolfe-Quintero et al. (1998), Ellis and Barkhuizen (2005), McNamara et al. (2014), Lu (2010, 2012), and Bulté and Housen (2012).

1.3 Linguistic complexity and SLA

Originating from Task-Based Language Teaching (TBLT) research where the construct has been predominantly used to account for task factors on learners' L2 performance (e.g. Ahmadian and Tavakoli, 2011; Crookes, 1989; Ellis and Yuan, 2004; Foster and Skehan, 1996; Yuan and Ellis, 2003), linguistic complexity has developed into a major variable for SLA research (Housen et al., 2009). Researchers have been analyzing the complexity of L2 learners' interlanguage for purposes such as (a) gauging proficiency, (b) describing performance, and (c) benchmarking development (Ortega, 2012). It has been used, on one hand, as dependent variables where effects of instruction, individual differences, learning context, and task design on the complexity of L2 performance can be investigated (Bytgate, 1996; Bygate, 1999; Collentine, 2004; De Graaff, 1997; Derwing and Rossiter, 2003; Foster and Skehan, 1996; Freed, 1995; Mora, 2006; Norris and Ortega, 2009). On the other hand, the construct has also been used as independent variables to evaluate language proficiency (Ortega, 2003) or measure longitudinal language development (Byrnes, 2009; De Clercq and Housen, 2017; Polat and Kim, 2014; Vyatkina et al., 2015). It is this latter use that makes linguistic complexity be considered as one of the primary foci of L2 research, rather than still being viewed as an interlanguage descriptor that results from factors affecting L2 acquisition (Housen et al., 2009; Towell, 2012).

Underlyingly, complexity is thought to have its own working mechanism, cognitive and psycholinguistic processes, and developmental dynamics (Larsen-Freeman, 2006; Robinson, 2011; Towell, 2012; Chen et al., Submitted-b).

Both theoretical account and empirical evidence have shown that linguistic complexity is closely related to L2 proficiency/development. Theoretically, SLA is postulated to consist of three basic needs: (a) the acquisition of an appropriate mental representation for 'linguistic competence' (Chomsky, 1986; Hawkins, 2001; White, 1990, 1991, 1992, 2003), (b) the acquisition of 'learned linguistic knowledge', and (c) the procedualization of the acquired knowledge (Levelt, 1989, 1999; Towell, 2012; Towell and Hawkins, 1994). These needs correspond to three types of learning: triggered, explicit, and procedural (Towell, 2012). Triggered learning is associated with the development of syntactic competence. Explicit learning makes it possible for L2 learners to purposely use more elaborate and varied language in situations where these characteristics are encouraged. Procedural learning has more to do with the fluent use of the learned knowledge. The three types of learning, the results of which are a comprehensive representation of the L2 knowledge in the learner's brain, enable learners to build up the ability to comprehend and produce the L2. As a result, the development of CAF can generally be seen as the 'product of successful interaction and integration between the growth of linguistic competence, the development of learned linguistic knowledge and the development of linguistic processing ability' (Towell, 2012, p. 66).

In particular, the complexity of L2 performance is largely determined by the learners' explicit interlanguage knowledge, such as their knowledge of the L2's lexis, grammar, and formulaic patterns, as well as how much this knowledge has been internalized and proceduralized (Housen et al., 2012). L2 development is associated with the learners' ability to control and make use of the ever-expanding repertoire of linguistic resources in their L2 (Foster and Skehan, 1996; Ortega, 2003, 2015). The results of this development is naturally more varied and elaborated language production, or higher complexity. However, adopting a developmental view of complexity does not mean that the ultimate goal of SLA is to produced increasingly more complex language as an end in itself. More complex language should not be automatically associated with higher proficiency or more development (Bulté and Housen, 2014; Pallotti, 2009, 2015).

A plethora of empirical research has also proved the connection between L2 complexity and proficiency/development. Studies have consistently shown that substantial exposure to and intensive instruction in the L2 would result in the increase of complexity measure values (Bulté and Housen, 2014; Crossley and McNamara, 2014;

De Clercq and Housen, 2017; Lu, 2011; Mazgutova and Kormos, 2015; Ortega, 2003; Ortega and Sinicrope, 2008; Vyatkina, 2012, 2013) in the students' L2 production. However, with regard to the areas of development or the developmental patterns, mixed even contradicting results have been found. For example, while Bulté and Housen (2014) found that their English as a Second Language (ESL) students significantly increased the lexical, but not the syntactic complexity of their L2 production within a four-month period, Vyatkina (2012) observed increased complexity in both areas.

In terms of the developmental patterns, both linear progression and the non-linear waxing and waning of linguistic complexity development have been observed (Bulté et al., 2008; Vyatkina, 2015; Larsen-Freeman, 2006). Multiple factors might have contributed to the different findings in previous research. One reason may be that different studies used different measures for the same complexity sub-constructs. As was reviewed in the previous section (Section 1.2), hundreds of complexity measures have been created to account for linguistic complexity. However, for most studies, usually only a few 'most representative' or 'best' measures were used to represent the complexity construct or its sub-constructs. This makes it hard to compare results from different studies and draw conclusions from contradicting findings.

In sum, linguistic complexity has proved to be a useful construct in SLA research. Not only has it been used to account for the traditional SLA problems such as input, individual differences, attrition, and output, it has also become one of the central foci of SLA inquiry. Findings from this research has provided new insights into SLA. However, as will be discussed in the next section and throughout this dissertation, there are still a number of gaps to be addressed for research on linguistic complexity.

1.4 Identifying the research gaps

As has been shown in the previous sections, there has been extensive research on the conceptualization (Section 1.1), measurement (Section 1.2), and application of complexity (Section 1.3). However, a lot of the findings in this research are still far from conclusive because of problems in the operationalization of complexity, the measurement instrument, and the nature of the studies, etc. It is thus difficult for L2 researchers and language education practitioners to apply the research results to the development of more effective teaching methodologies or to their daily teaching practice. Firstly, insufficient research has been conducted to address the multidimensionality of the complexity construct. While they acknowledged complexity as a multifaceted, multilayered construct, most previous studies still adopted a reduc-

tionistic approach because of the lack of efficient instrument to help them measure the multiple dimensions. Developmentally, controversies on the relationship between complexity and the other two constructs in the CAF triad continue to baffle SLA researchers. Furthermore, there is a general lack of intervention studies in the research on complexity development. Most of the studies up to date are observational in nature, making it difficult to draw causal conclusions on certain phenomena. It would be optimal if SLA researchers could figure out how to make use of the complexity construct to help learners better promote their language proficiency. Such discovery could only be made with the help of intervention studies. Last but not least, the great potential of complexity analysis to be implemented in ICALL systems for promoting L2 development has not been fully exploited. Linguistic complexity analysis is highly automatizable with modern NLP technologies, making it an optimal candidate to model both L2 input and production. But surprisingly, practical ICALL systems implementing the complexity research findings are still scarce.

This section is devoted to a more detailed discussion of these four gaps from previous complexity research. The next section will then provide an overview of the work we have done to try to fill these gaps. Detailed reports of the projects are presented in the rest of the chapters.

1.4.1 Lack of comprehensive tools for complexity analysis

The proliferation of linguistic complexity measures is observable from a number of studies (see, for example, Wolfe-Quintero et al., 1998; Ellis and Barkhuizen, 2005; McNamara et al., 2014, for inventories of complexity measures). It has become extremely difficult to extract all these complexity values by hand. However, projects adopting a data-driven approach are required to explore a large number of measures from various linguistic aspects in order to have a more complete representation of the construct. As a result, automatic tools geared towards linguistic complexity analysis have emerged in the past few years. Examples of these tools and analysis platforms include Xiaofei Lu's Syntactic and Lexical Complexity Analyzers[1] (Lu, 2010, 2012), CohMetrix[2] (McNamara et al., 2014), Kristopher Kyle's Suite of Linguistic Analysis Tools[3] and so on. These tools are efficient for extracting complexity measures of certain linguistic aspects, such as lexical, syntactic, or discoursal aspects, but one needs to go through all these tools to obtain a comprehensive set of the complexity measures. These analysis platforms were developed with different programming

[1] `http://www.personal.psu.edu/xxl13/download.html`
[2] `http://www.cohmetrix.com/`
[3] `http://www.kristopherkyle.com/tools.html`

languages and provide different user interfaces, making them challenging to use by non-expert computer users. Furthermore, few systems support collaborative development, causing duplications of measures across systems and waste of research resources. As a result, it would be optimal if any automatic system could offer a one-stop solution to comprehensive complexity analysis to researchers interested in using the construct for various research purposes. From the user's perspective, the system should also be easy to use for those with little programming experience. Additional functionalities to help the user manage corpora, select complexity measures, and visually explore the analysis results would also be highly welcome and demanded.

1.4.2 Unclear developmental interrelationship among CAF

There has been a plethora of cross-sectional and longitudinal research on the development of both complexity and accuracy (e.g. Foster and Skehan, 1996; Ortega, 2003, 2015; Vyatkina, 2012; Bulté and Housen, 2014; Vyatkina et al., 2015; Yoon and Polio, 2017; Larsen-Freeman, 2006; Verspoor et al., 2012), as well as on the interrelationship between the two constructs (e.g. Skehan, 1998; Robinson, 2001, 2005; Vercellotti, 2017; Yuan and Ellis, 2003). The latter inquiry is especially interesting to SLA researchers and practitioners because of its potential impact on L2 teaching practice. If trade-off between complexity and accuracy exists due to the limited attentional resources of the learners and the competition for such limited resources when they try to produce the L2, as was predicted by some researchers (VanPatten, 1990; Kuiken and Vedder, 2007; McCutchen, 1996; Skehan, 1998, 2009), teachers might give priority to either one aspect in L2 instruction. On the contrary, if the interrelationship between complexity and accuracy can be mutually promotional, as was hypothesized by Robinson's (2001; 2003) Cognition Hypothesis (CH), it would then be possible to design tasks that help promote the simultaneous development of both L2 accuracy and complexity (Robinson and Gilabert, 2007). The controversy in this research is still unsettled. Both Skehan's Trade-Off Hypothesis (TOH) and Robinson's CH found empirical support from TBLT experiments, which varied task factors such as planned *vs.* unplanned, or monologic *vs.* dialogic and investigated their effects on CAF. As a result, the contradicting findings and the limited experimental contexts in which the findings are discovered make it still unclear about the interrelationship between the CAF constructs. Second language researchers and teaching practitioners would benefit from a more clear answer to this question. It would be much more convincing if such an answer could also be supported by longitudinal data of L2 development.

1.4.3 Lack of intervention studies linking input and production complexity

Linguistic complexity has been used as both dependent and independent variables in SLA studies (see Section 1.3) to study task effects on language production, to evaluate language proficiency, and to measure longitudinal language development (Ortega, 2012). It can be used to analyze both learning input and learner production (Meurers, 2012). Research analyzing the two types of language production separately has yielded fruitful results. For analyzing learning input, complexity has been successfully used for readability assessment (Collins-Thompson, 2014; Vajjala and Meurers, 2012), writing quality evaluation (Ferris, 1994; Taguchi et al., 2013). The analysis of learner production complexity has also been applied to characterize performance of learners of different developmental stages (Bulté and Housen, 2018; Vyatkina, 2012, 2013), from different L1 backgrounds (Lu and Ai, 2015), and/or under different task contexts (Alexopoulou et al., 2017). What is still lacking from this line of research is the unification of complexity from the both spaces: input and production. It is well-acknowledged in SLA that there is an effect of input on learner production or proficiency development (e.g., the Input Hypothesis (IH), Krashen, 1985). It is also well-established that with increasing L2 proficiency, learners usually exhibit the ability to make use of a wider range of linguistic resources and more elaborate language to express their ideas, making the produced language more complex (Jarvis et al., 2003; Laufer and Nation, 1995; Verspoor et al., 2008; Lu, 2011). However, little has been done to explore the effects of complex input on L2 proficiency development. The observational nature of previous research makes it difficult to tease apart the causes of proficiency development from the perspective of linguistic complexity. It is thus difficult for L2 instructors and ICALL developers to select teaching materials or design tutoring systems for promoting L2 proficiency by making use of the complexity construct. Consequently, there arises a need for an intervention study experimenting on the link between complex input and proficiency development.

1.4.4 Scarcity of ICALL systems implementing complexity research findings

The operationalization of complexity as the variedness and elaborateness of language production (Ellis, 2003) makes the construct quantifiable and computationally automatizable. Although a few analytical systems for extracting complexity measures have been developed (see Section 1.4.1), few ICALL systems making use of the re-

search findings on linguistic complexity have been developed. Several reasons might have contributed to the scarcity of ICALL systems in this respect. Firstly, automatic analysis of linguistic complexity is a recent development. Most analytical tools were developed in the past few years, following the maturity of modern NLP tools. Previous research on complexity analysis has mainly focused on the validation of these tools. Secondly, as was discussed in Section 1.4.3, the causal effect of complex input has not been confirmed with intervention experiments. Effective design of ICALL systems will need to be based on the findings from such experiments. Last but not least, successful implementation of ICALL systems requires knowledge of both SLA and computer technology, which entails collaboration between SLA researchers and computer science experts. Unfortunately, such collaboration is yet to become the norm of the two fields. As a result, although there is a great potential of complexity in ICALL system design, more needs to be done for the actual implementation of such systems.

1.5 Overview of the dissertation

The rest of this dissertation will focus on addressing the research gaps identified in the previous section. Specifically, Chapter 2 addresses the reductionistic view of complexity (Section 1.4.1) by presenting a system supporting the comprehensive analysis of linguistic complexity—the CTAP. The system features a friendly and easy-to-use interface, a collaborative developmental framework, additional functionalities for corpus management, flexible complexity feature selection, and results data visualization.

Chapter 3 investigates how lexical sophistication measures utilizing word frequency lists can be applied to text readability assessment. It will show how a low-level complexity measure of lexical sophistication can be made full use of to characterize a high-level construct of text readability. The study also shows how the multidimensionality of complexity can be addressed more fully from a micro- and individual-dimensional level.

Chapter 4 tries to tackle the developmental interrelationship between complexity and accuracy (Section 1.4.2), an inquiry that interests a lot of SLA researchers and practitioners. The data used in this study were collected longitudinally under natural instructional settings, which are different from most previous studies where the data were mainly from teaching experiments, making our findings more convincing and relevant to actual teaching practice.

Chapter 5 tries to connect the complexity of learning input and learner produc-

tion by calculating the distance between complexity feature vectors comprising of multiple complexity dimensions representing the input and production. The study confirms the validity of vector distance as a link between input and production complexity. An ICALL system implementing the findings is reported in Chapter 6, which also addresses the ICALL system scarcity issue discussed in Section 1.4.4.

With the link between input and production created with the complexity construct, Chapter 7 further tests the effects of complex input on the development of the learners' L2 proficiency with an intervention study to address the lack of intervention research (Section 1.4.3).

The dissertation concludes with Chapter 8, which summarizes the main findings of the studies reported in this dissertation, as well as points to further directions for complexity research.

Chapter 2

Automatic multidimensional analysis of complexity—The Common Text Analysis Platform

Chapter highlights

What is already known about this topic:

- Multidimensional analysis of linguistic complexity has been widely applied to multiple fields of language education research.
- Sophisticated tools have been developed by computational linguists for the automatic analysis of linguistic complexity.

What this study adds:

- We developed a text analysis system trying to create the connection between the features that can in principle be identified based on state-of-the-art computational linguistic analysis, and the analyses a teacher, textbook writer, or second language acquisition researcher can readily obtain and visualize for their own collection of texts.
- The system supports fully configurable linguistic feature extraction for a wide range of complexity analyses and features a friendly user interface, integration of modularized and reusable analysis components, and flexible corpus and feature management.

Implications for theory, policy, or practice:

- For theory: The CTAP system makes it easier for SLA researchers and language teaching practitioners to address the multidimensionality of the complexity construct, hence avoiding a reductionistic view of it.

- For practice: We created a common platform for complexity analysis, encouraging research collaboration and sharing of feature extraction components to jointly advance the state-of-the-art in complexity analysis in a form that readily supports real-life use by ordinary users.

Abstract

Informed by research on readability and language acquisition, computational linguists have developed sophisticated tools for the analysis of linguistic complexity. While some tools are starting to become accessible on the web, there still is a disconnect between the features that can in principle be identified based on state-of-the-art computational linguistic analysis, and the analyses a teacher, textbook writer, or second language acquisition researcher can readily obtain and visualize for their own collection of texts.

This chapter presents a web-based tool development that aims to meet this challenge. The Common Text Analysis Platform (CTAP) is designed to support fully configurable linguistic feature extraction for a wide range of complexity analyses. It features a user-friendly interface, modularized and reusable analysis component integration, and flexible corpus and feature management. Building on the Unstructured Information MAnagement (UIMA) framework, CTAP readily supports integration of state-of-the-art NLP and complexity feature extraction maintaining modularization and reusability. CTAP thereby aims at providing a common platform for complexity analysis, encouraging research collaboration and sharing of feature extraction components to jointly advance the state-of-the-art in complexity analysis in a form that readily supports real-life use by ordinary users.

Related publication

This chapter is based on the following publication:

- Chen, X. and Meurers, D. (2016b). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity at COLING*, pages 113–119, Os-

aka, Japan, 11th December. The International Committee on Computational Linguistics.

2.1 Introduction

Linguistic complexity is a multifaceted construct used in a range of contexts, including the analysis of text readability, modeling the processing difficulty of sentences in human sentence processing, analyzing the writing of second language learners to determine their language proficiency, or for typological comparison of languages and their historical development. To analyze linguistic complexity in any of these contexts, one needs to identify the observable variedness and elaborateness (Rescher, 1998; Ellis, 2003) of a text, which can then be interpreted in relation to the nature of the task for which a text is read or written, or the characteristics of the individuals engaged in reading or writing. This chapter mainly concerns about the first step: identifying the elaborateness and variedness of a text, sometimes referred to as absolute complexity (Kusters, 2008, cf. Section 1.1), or linguistic complexity.

Measure of absolute complexity for the purpose of selecting reading materials or the analysis of learner language ranges from more holistic, qualitative perspectives to more analytic, quantitative approaches. While we here focus on the latter, reviews of both can be found in Pearson and Hiebert (2014); Collins-Thompson (2014); Benjamin (2012); Ellis and Barkhuizen (2005), and Wolfe-Quintero et al. (1998). The quantitative measurement of complexity can be done on all levels of linguistic representations (lexical, morphological, syntactic, and phonological, cf. Section 1.1), making the number of complexity measures so large that it is difficult, if not impossible, to extract all these measures from corpora of non-trivial sizes.

This chapter describes a system that supports the extraction of quantitative linguistic features for absolute complexity analysis: the Common Text Analysis Platform (CTAP). CTAP is an ongoing project that aims at developing a user-friendly environment for automatic complexity feature extraction and visualization. Its fully modularized framework enables flexible use of NLP technologies for a broad range of analysis needs and collaborative research. In the following sections, we first sketch demands that a system for complexity analysis and research should satisfy, before providing a brief description of the CTAP modules and how they are integrated to address the demands.

2.2 Identifying demands

In order to find out how complexity had been measured in L2 research Bulté and Housen (2012) reviewed forty empirical studies published between 1995 and 2008 and compiled an inventory of 40 complexity measures used in these studies (pp. 30–31). Although they found that there was ‘no shortage of complexity measures in

SLA studies', most studies used no more than 3 indices to measure complexity. This was largely 'due to the lack of adequate computational tools for automatic complexity measurement and the labor-intensiveness of manual computation' (p. 34). The authors were optimistic that some online complexity analyzers would come out in the near future and the situation would change.

As Bulté and Housen predicted, a number of complexity analysis tools were released in the past few years (e.g., Xiaofei Lu's SCA and LCA[1], CohMetrix's Web interface to its 106 complexity features[2], and Kristopher Kyle's Suite of Linguistic Analysis Tools[3], etc.). While they make it possible for researchers to measure absolute linguistic complexity easier and faster, these tools were generally not designed for collaborative research and are limited in terms of usability and platform compatibility, provide no or very limited flexibility in feature management, and do not envisage analysis component reusability. As a result, they are not suitable (and generally were not intended) as basis for collaborative research on complexity, such as joint complexity feature development.

Commercial systems such as the TextEvaluator[4] by the Educational Testing Service (ETS) and the Reading Maturity Metric[5] by Pearson Education also implemented automatic complexity analysis for readability assessment (See Nelson et al., 2012, for a comprehensive review and assessment of such systems.) However, the commercial nature of these systems limits the transparency of the mechanisms they employ and future research cannot be freely developed on this basis. The Text Analysis, Crawling, and Interpretation Tool (TACIT, Dehghani et al., 2016) provides an open-source platform for text analysis. While linguistic complexity analyses could be integrated in this framework, it so far is primarily geared towards crawling and text analysis in a social media context, e.g., for sentiment analysis.

These complexity analysis tools overlap in terms of the complexity features offered by different systems. For example, the tools exemplified earlier contain a significant amount of lexical feature overlap across systems. While this can be useful for cross-validating the calculated results, it also duplicates analyses options without giving the user the choice of selecting the set of analyses needed to address the specific needs. A more optimal scenario would be based on a common framework where developers of feature extraction tools can collaborate and share analysis components, release analysis tools to be used by researchers who focus on different

[1] http://www.personal.psu.edu/xxl13/download.html

[2] http://cohmetrix.com

[3] http://www.kristopherkyle.com

[4] Formerly SourceRater, cf. https://texteval-pilot.ets.org/TextEvaluator

[5] http://www.pearsonassessments.com/automatedlanguageassessment/products/100000021/reading-maturity-metric-rmm.html#tab-details

aspects of the complexity problems (e.g., relative complexity for a specific target audience).

Another issue of existing complexity analysis tools concerns (re)usability. Many of these tools are released as standalone pre-compiled software packages or program source code. Pre-compiled packages not only cause cross-platform compatibility problems, but also are difficult to adapt to meet the user's specific needs. The source code option provides maximum flexibility, but are usable only to expert users or programmers. It should be noted that a lot of complexity researchers are linguists, psychologists, or cognitive scientists, but not necessarily computer scientists or programmers. Consequently, developing a complexity analysis system with user-friendly interface and visualization features are on demand.

Last but not least, there is also the challenge of complexity feature proliferation over the past years. Researchers are systematically exploring and identifying new features that contribute to our understanding of linguistic complexity. For example, CohMetrix (McNamara et al., 2014) provides 106 metrics for measuring cohesion and coherence. Housen (2015) identified more than 200 features for measuring L2 complexity. Vajjala (2015) accumulated another 200 features for doing readability assessment. Although features overlap across systems, the number of complexity features used and compared by researchers is large and likely to grow . Not every study needs to use all these features, nor any tool provides a full set. Researchers interested in linguistic complexity arguably would benefit from a system that readily supports them in choosing and applying complexity analyses from a large repository of features, without requiring NLP expertise.

2.3 System architecture of CTAP

The CTAP system is designed to address the issues reviewed in the previous section. The goal is a system that supports complexity analysis in an easy-to-use, platform independent, flexible and extensible environment. The system consists of four major user modules—Corpus Manager, Feature Selector, Analysis Generator, and Result Visualizer—as well as a Feature Importer administrative module. Figure 2.1 shows the system architecture and module relationships.

The Corpus Manager helps users manage the language materials that need to be analyzed. They can create corpora to hold texts, folders to group corpora and tags to label specific texts. The text labels will then be used to help filter and select target texts for analysis. They can also be used to group texts for result visualization purposes.

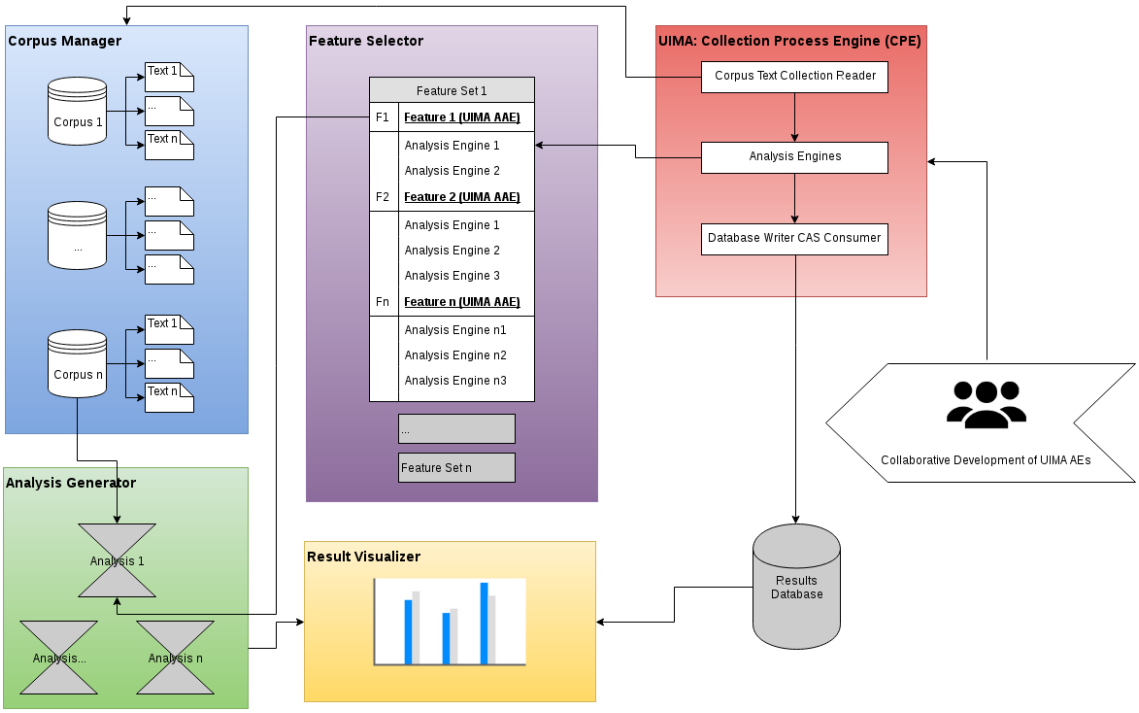

Figure 2.1: CTAP modules and their relationship

Other complexity analyzers usually limit users to a fixed set of features that the analyzer extracts. The Feature Selector from CTAP enables users to group their selection of the complexity features into feature sets. This flexibility is realized by utilizing the UIMA framework[6] provided by the Apache Foundation. By using the UIMA framework, every complexity feature can be implemented as an Aggregate Analysis Engine (AAE) which chains up a series of primitive Analysis Engines (AEs). Each AE may be a general purpose NLP components, such as a sentence segmenter, parser, or POS tagger. It may also be one that calculates some complexity feature values based on analysis results from upstream AEs or components. This setup enables and encourages reusability of AEs or analysis components, thus making collaborative development of complexity feature extractors easier and faster.

After collecting/importing the corpora and selecting the complexity features, the users can then generate analyses in CTAP's Analysis Generator. Each analysis extracts a set of features from the designated corpus. Results of the analysis are then persisted into the system database and may be downloaded to the user's local machine for further processing. The user can also choose to explore analysis results with CTAP's Result Visualizer. The UIMA framework supports parallel computing that can easily scale out for handling big data analysis needs.

The Result Visualizer is a simple and intuitive module that plots analysis results

[6] `https://uima.apache.org`

for the user to visualize preliminary findings from the analysis. It supports basic plot manipulation and download. Figures 2.2–2.5 show screenshots of the user modules.

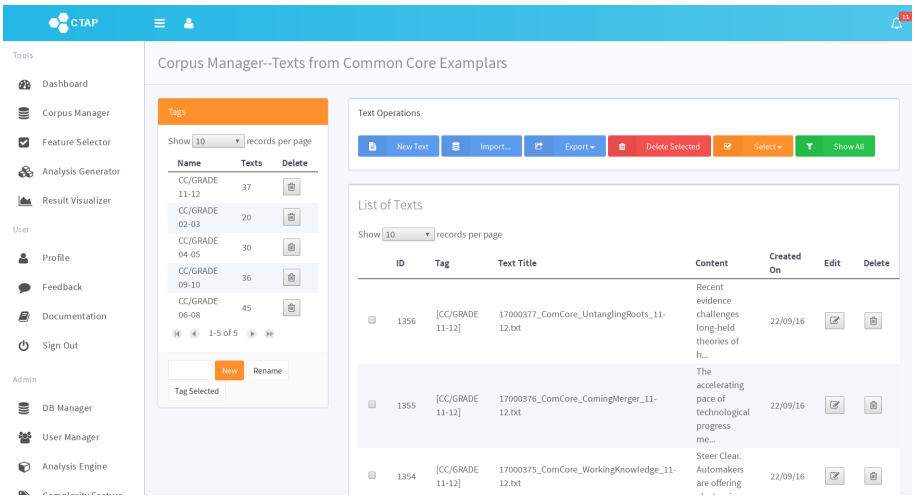

Figure 2.2: Corpus Manager module screen shot

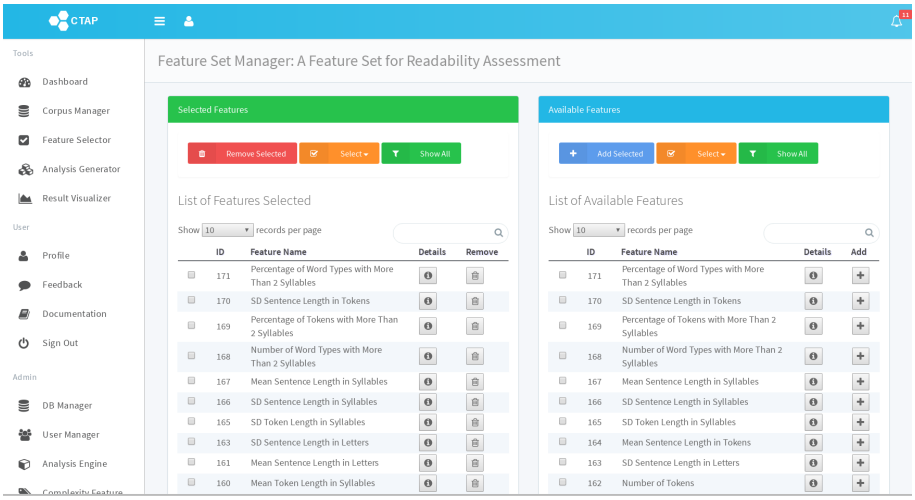

Figure 2.3: Feature Selector module screen shot

2.4 Design features of CTAP

The target users of the CTAP system are complexity feature developers and linguists or psychologists who might not necessarily be computer science experts. As a result, the system features the following design.

Consistent, easy-to-use, friendly user interface. The CTAP system is deployed as a Web application, which strikes a balance between usability, flexibility

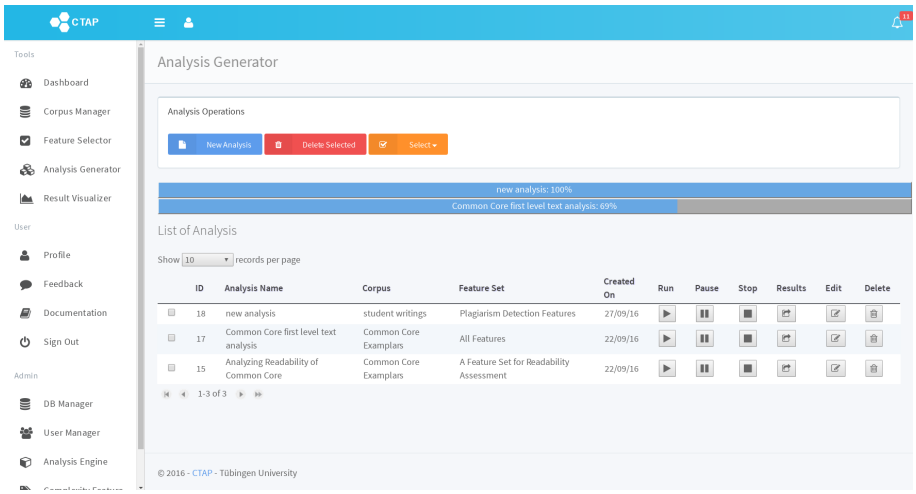

Figure 2.4: Analysis Generator module screen shot

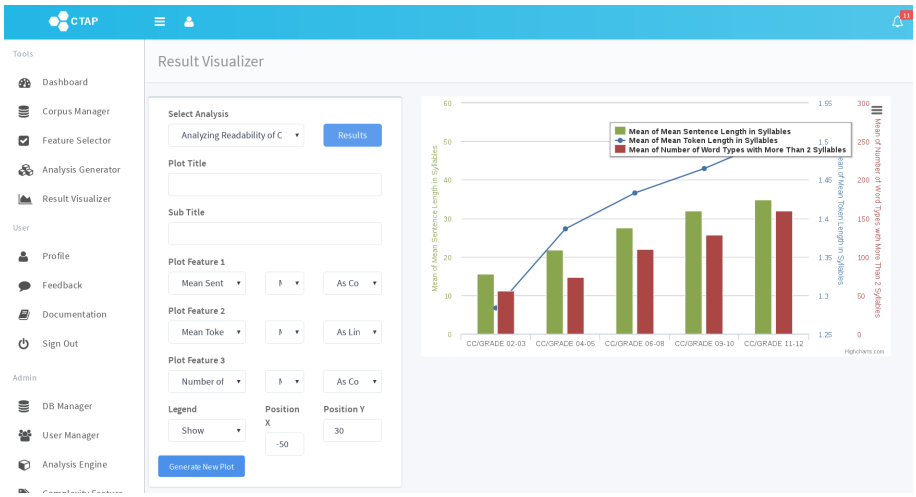

Figure 2.5: Result Visualizer module screen shot

and cross-platform compatibility. The Graphical User Interface (GUI) provided on the Web makes it easy to access, user-friendly and platform neutral. The CTAP client frontend was written with Google Web Toolkit (GWT)[7], an open source and free technology that enables productive development of high-performance web applications. This avoids the necessity to compile the software for different operating systems, which has been proved to be a major frustration for small development teams or single developers who do not have enough resources to deal with platform differences.

Modularized, reusable, and collaborative development of analysis components. The CTAP analysis back-end is written under the UIMA framework. Each

[7] `http://www.gwtproject.org`

analysis unit is implemented as a UIMA AE. Since a lot of the AEs are commonly required by different complexity features, modularizing analysis into smaller AEs makes it easier to reuse and share components. The AEs included into CTAP are open sourced and we encourage contribution from feature developers. A community effort will enhance complexity research to a greater extent.

Flexible corpus and feature management. This feature is a luxury in light of the existing complexity analysis tools. However, this feature is of special value to users with lower information and communication technology competence. Users choose from the feature repository the system provides a set of features that meet their needs, the CTAP system then generates a UIMA AAE to extract the chosen feature values. It frees users from tediously editing analyzer source code, which is also often error-prone.

2.5 System and source code availability

The CTAP project is under active development at the moment. A demo version of the system has been finished and made available at `http://www.ctapweb.com`, establishing the feasibility of the design, architecture, and the features described in this chapter. The current collection of complexity measures in the system contains over 170 indexes. The addition of all the measures listed in Appendix A is well underway. Since the architecture of the CTAP system is language independent, new feature extractors supporting other languages can be easily plugged into the system as UIMA AEs. A component supporting analysis of German texts is being developed by our colleagues who work on German complexity analysis.

In making the tool freely available under a standard Creative Commons by-nc-sa license, we would also like to call for contribution from other researchers. Interested parties are encouraged to join and contribute to the project at `https://github.com/ctapweb`. Only by making use of joint effort and expertise can we envisage a production level system that can support joint progress in the complexity research community, while at the same time making the analyses readily available to ordinary users seeking to analyze their language material—be it to study language development or to develop books better suited to the target audience.

2.6 Summary

The analysis of linguistic complexity of learning input and learner production is applicable to a lot of SLA research. The multidimensionality of complexity results in the proliferation of complexity measures for different research purposes. There is still a general lack of comprehensive tools for automatic complexity analysis (see review in Section 1.4.1) to help L2 researchers to approach the multidimensionality of complexity, especially for those who are not familiar with NLP technologies. The CTAP system provides a solution to this problem by creating a general framework to streamline the analytical process. As a result, it helps to fill the gap of the lack of comprehensive analytical tools to tackle the multidimensionality of the complexity construct. As will be seen in the following chapters, the CTAP tools provide researchers with a convenient method to investigate interesting SLA issues from the complexity perspective.

Chapter 3

Linguistic Complexity and Readability Assessment

Chapter highlights

What is already known about this topic:

- Linguistic complexity has been widely used for readability assessment. The lexical sophistication measures utilizing word frequencies have been found to be highly predictive of text readability.
- The effectiveness of word frequency as a predictor of text readability is based on the cognitive model that frequently used words have a higher base level of activation and consequently require less additional activation for retrieval from the reader's mental lexicon.
- Based on the frequencies of words in corpora assumed to be representative of language experience, readability research commonly uses the mean frequencies of all the words in a document to characterize its readability.

What this study adds:

- The study investigates the impact of frequency norms derived from different corpora on readability assessment in different testing setups.
- It compares different types of frequency measures, from occurrence counts to counts of the number of contexts in which a word is used as well as their normalized variants, for readability assessment.
- It explores three approaches to characterize text-level readability from the word-level complexity measure of lexical sophistication: from the standard

deviation of the word frequencies in a document, via the mean frequencies of the words in particular language frequency bands, to the mean frequencies of the document's words grouped by agglomerative clustering.

Implications for theory, policy, or practice:

- For theory: Lexical frequency can be highly predictive of text level readability, in line with the cognitive model of word frequency effects on reading.

- For practice: High quality readability assessment depends on well-chosen reference corpora and a method for aggregating lexical frequency information that represents the distribution of word frequencies in a text more richly than using a single mean.

Abstract

Assessment of text readability is important for assigning texts at the appropriate level to readers at different proficiency levels. The present research approached readability assessment from the lexical perspective of word frequencies derived from corpora assumed to reflect typical language experience. Three studies were conducted to test how the word-level feature of word frequency can be aggregated to characterize text-level readability. The results show that an effective use of word frequency for text readability assessment should take a range of characteristics of the distribution of word frequencies into account. For characterizing text readability, taking into account the standard deviation in addition to the mean word frequencies already significantly increases results. The best results are obtained using the mean frequencies of the words in language frequency bands or in bands obtained by agglomerative clustering of the word frequencies in the documents—though a comparison of within-corpus and cross-corpus results shows the limited generalizability of using high numbers of fine-grained frequency bands. Overall, the study advances our understanding of the relationship between word frequency and text readability and provides concrete options for more effectively making use of lexical frequency information in practice.

Related publications

This chapter is based on the following publications:

- Chen, X. and Meurers, D. (2018b). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

- Chen, X. and Meurers, D. (2016a). Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications at NAACL*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

3.1 Introduction

Successful reading comprehension depends, to a large extent, on how well teachers or students are able to select materials that match the students' reading abilities. Reading materials that match with students' reading ability provide them with useful practice and self-learning, but are not too hard to make them feel frustrated (Chall et al., 1991). Students usually gain a sense of success and are motivated to read more when they are given texts that enable them to practice being competent readers (Milone and Biemiller, 2014). As a result, it is very important that teachers and students are equipped with tools or methods to evaluate the readability of reading materials, which is defined as the sum of all elements of a text that affects a reader's understanding, reading speed, and level of interest in the text (Dale and Chall, 1949). However, despite a research history of nearly a century (see Zakaluk and Samuels, 1988; Benjamin, 2012; Collins-Thompson, 2014, for reviews of this research), readability assessment still poses a challenge not only to students, but also to researchers and language teachers.

Early research on readability focused on the construction of multiple-regression formulas for predicting the reading levels of texts with some surface semantic and syntactic features such as sentence length and average word length in syllables (e.g. Dale and Chall, 1948; Flesch, 1948; Gray and Leary, 1935; Kincaid et al., 1975; McLaughlin, 1969; Vogel and Washburne, 1928). Later research looked at deeper structural and cognitive variables such as propositional density and coherence for predicting text readability (e.g. Crossley et al., 2008; Graesser et al., 2004; Kintsch et al., 1993; McNamara et al., 2010). Recent research has focused on the separate and combined effects of lexical (Crossley et al., 2007; Flor et al., 2013; Lu et al., 2014), morphological (François and Watrin, 2011; Hancke et al., 2012), psycholinguistic (Boston et al., 2008), semantic (vor der Brück et al., 2008), syntactic (Heilman et al., 2007), and cognitive (Feng, 2010; Feng et al., 2009; Flor and Klebanov, 2014; Foltz et al., 1998; Graesser et al., 2011; Wolfe-Quintero et al., 1998) features on readability by making use of the latest development in NLP technologies and Machine Learning (ML) methods. Although more and more linguistic and cognitive features have been incorporated into the readability assessment models, it was found that the semantic variable of word difficulty accounts for the greatest percentage of readability variance (Marks et al., 1974).

One way of measuring word difficulty is to use the frequency of the words calculated from a corpus of that language's general use (Ryder and Slater, 1988), a frequency norm of the language. The cognitive basis of frequency norms as a proxy to word difficulty is the finding that high-frequency words are more easily

perceived (Bricker and Chapanis, 1953) and readily retrieved (Haseley, 1957) by language users, making them 'easier' than low-frequency ones. As a result, a frequency norm that faithfully represents the language users' exposure and experience with the language would be predictive of word difficulty as perceived by the language users. Previous research has shown the effectiveness of using word frequency norms for readability assessment (Lexile, 2007; Milone and Biemiller, 2014; Ojemann, 1934; Patty and Painter, 1931). However, besides an over-simplifying use of frequency norms, little research has probed into the nature of the frequency norms, the frequency measures most appropriate for readability assessment purposes, or how the measures can be better used to improve predictive accuracy.

The present study tries to extend readability research from the lexical perspective. Our interest is in the use of word frequency norms for readability assessment, an issue that had caught on since the very beginning of readability research but yet to be settled. This study merits itself not only in providing a better understanding of the relationship between word frequency and text readability, but also in pointing to new methods on how to better aggregate a word-level feature of a text to predict the text-level characterization of text readability.

In the following sections, we will first discuss how and why word frequency is related to reading comprehension and review how lexical complexity or frequency variables had been used in earlier readability studies. The review will help identify the need for further inquiry into the relationship between vocabulary frequency and readability. Followed are the descriptions of the experiments we ran and their results. New insights into how to characterize textual difficulty with frequency norms will then be reported and discussed.

3.2 Reading comprehension and the word frequency effects

Reading is viewed as a coordinated execution of a series of processes, including word encoding, lexical access, assigning semantic roles, and relating the information contained in a sentence to earlier sentences in the same text and the reader's prior knowledge (Just and Carpenter, 1980). These processes require that the readers possess the corresponding grammatical and syntactic skills necessary for decoding sentences in the text. In addition to syntactic competence, the reader's semantic decoding abilities also play an important role in successful reading comprehension (Marks et al., 1974). Understanding of a text begins with relating the print words to the vocabulary the reader previously acquired. The connection thus cre-

ated enables the reader to draw from previous experience meanings and concepts to make sense of the reading text. In order for this to happen, the reader must have a sufficient mastery of the vocabulary in the language with which the text is written.

Vocabulary knowledge has been proved vital to reading comprehension (Laufer and Ravenhorst-Kalovski, 2010; Nation, 2006). Laufer and Ravenhorst-Kalovski (2010) examined the relationship between lexical text coverage, learners' vocabulary size and reading comprehension. They found that even a small increment of vocabulary knowledge would result in sizable increase in reading comprehension. One of their conclusions was that the lower lexical coverage of frequent words was a characteristic of difficult text, while high lexical coverage from frequent words made texts easier to understand. An important implication from this research is that factors such as lexical coverage and vocabulary knowledge are good predictors of reading comprehension, an idea shared by a number of other researchers (e.g. Bernhardt and Kamil, 1995; Laufer, 1992; Nation, 2001, 2006; Qian, 1999, 2002; Ulijn and Strother, 1990).

A reader's vocabulary knowledge is related to the amount of exposure the reader has received on words. The more a word appears in various contexts, the more likely it is to be met and acquired by the reader. Word frequency is predictive to word difficulty (Ryder and Slater, 1988). Leroy and Kauchak (2014) evaluated the relationship between a reader's familiarity with a word and the word's frequency in common English text. They found that word frequency is strongly associated with both actual difficulty (how well people can choose the correct definition of a word) and perceived difficulty (how difficult a word looks). In general, high-frequency words are more easily perceived (Bricker and Chapanis, 1953) and readily retrieved by the reader (Haseley, 1957). High-frequency words are perceived and produced more quickly and more efficiently than low-frequency ones (Balota and Chumbley, 1984; Howes and Solomon, 1951; Jescheniak and Levelt, 1994; Monsell et al., 1989; Rayner and Duffy, 1986), resulting in more efficient comprehension of the text (Klare, 1968). Quoting Johnson et al. (1960) and Klare et al. (1955), who found a close relationship between frequency and students' reading preference, Klare (1968) concluded that the frequency of occurrence of words affects not only the ease of reading, but also its acceptability. We consider these as the frequency effects of vocabulary on reading comprehension, as illustrated in Figure 3.1. The frequency effects are based on the cognitive model that frequently used words have a higher base level of activation, and consequently require relatively less additional activation while being retrieved from the reader's mental lexicon (Just and Carpenter, 1980). Just and Carpenter (1980) validated this hypothesis with eye-tracking ex-

periments, in which they found a strong correlation between the frequency measures from the Kučera and Francis Frequency List (Kučera and Francis, 1967) and the gaze durations of words by the readers. Three types of psychological mechanisms underlie the frequency effects of language acquisition, comprehension and production: the strengthening of linguistic representations, the strengthening of linguistic expectations, and the development of automatized chunks (Diessel, 2007).

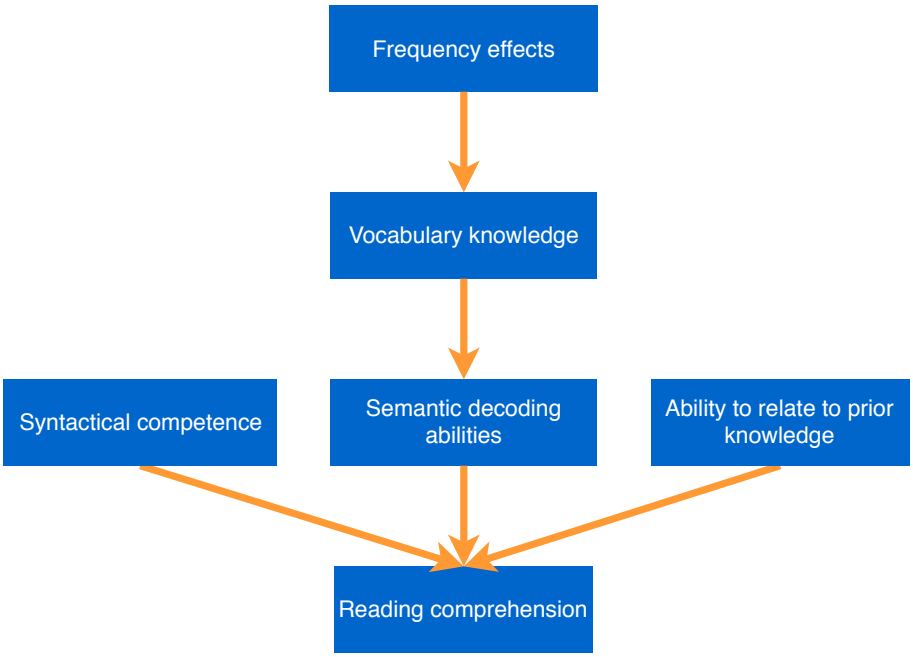

Figure 3.1: The frequency effects of vocabulary on reading comprehension.

3.3 Readability assessment with word frequency

Based on findings that relate reading comprehension to vocabulary knowledge and the latter to word frequency of occurrence, it is consequently reasonable to believe that vocabulary frequency is a good predictor of reading difficulty or text readability[1]. In the field of readability assessment, researchers have constantly used semantic and syntactic features of a text to predict its difficulty level (Dale and Chall, 1948; Flesch, 1948; Gray and Leary, 1935; Kincaid et al., 1975; Kintsch and Vipond, 1979; Kintsch et al., 1993; Lexile, 2007; Vajjala and Meurers, 2012). It was found that the semantic variable of word difficulty usually accounted for the greatest percentage of readability variance (Marks et al., 1974). As reviewed earlier, word

[1]We do not distinguish between text difficulty and text readability in this study, although some researchers consider them differently. Text difficulty is sometimes viewed as relative to the readers' reading ability, background knowledge or motivation. The terms text difficulty and readability in this study both refer to the 'absolute' sense of text readability.

difficulty is predictable by word frequency of occurrence. Consequently, we believe that the readability of a text is assessable by investigating the frequency of the words chosen for writing the text. This view is further supported by the contention that lexical frequency and diversity play an important role in reading difficulty or comprehension (Laufer and Ravenhorst-Kalovski, 2010; Marks et al., 1974; Nation, 2006; Schmitt et al., 2011).

Research on word frequency for text readability assessment abounds. It can be dated back to the earliest readability studies. For example, two frequency-related variables were tested by Lively and Pressey (1923) for creating readability formula: the number of 'zero-index words' and the median of the index numbers of words from Thorndike's list of the 10,000 most frequent words in English—*The Teacher's Word Book* (Thorndike, 1921). Lively and Pressey found that the median index number was the best indicator of the vocabulary burden of reading materials. *The Teacher's Word Book* was also referenced in other readability formula studies, such as Patty and Painter (1931) and Ojemann (1934). Patty and Painter (1931) calculated the average word weighted value, which was the average of the products of the index value from Thorndike's list and the frequency of the words in the text sample, and found 'an apparent improvement in technique' for readability judgment. Ojemann (1934) found the percent of words from a text that are among the first 1,000 and first 2,000 most frequent words of the Thorndike list highly correlated with difficulty. Word frequency measures are also used by the latest commercial readability products—Lexile (2007) and ATOS (Milone and Biemiller, 2014), which are not only commercially successful, but have also been proved effective (Nelson et al., 2012). The Lexile Framework (Lennon and Burdick, 2014) makes use of word frequencies from the Carrol-Davies-Richman corpus (Carroll et al., 1971) for their formula, while ATOS uses the Graded Vocabulary list.

A common problem with this research is that while investigating the regressional correlation between the frequency variables and the texts' reading levels, little attention has been paid to the characteristics of the frequency lists themselves. The review from the previous section reveals the connection between frequency and vocabulary difficulty, which in turn influences reading comprehension or difficulty. But this connection should be based on the fact that the frequency norm is a faithful representation of the reader's actual language experience. If it does not reflect how often, or in how many different situations the reader has encountered a word, it will not be able to predict the ease of retrieval or perception accurately. Hence, it will not be a good predictor of reading comprehension. Frequency lists such as *The Teacher's Word Book* and the Carroll-Davies-Richman (Carroll et al., 1971) list

used by early readability formulas were based on written corpora. Although these corpora were carefully constructed from materials that students read in their daily life, they failed to represent the also important, if not more important, source of spoken language that students are exposed to. The amount of exposure from spoken language is much greater than from written ones. Failure to include the spoken language does great harm to the representativeness of the frequency list and to its predictive power for text readability.

The measures used in these frequency lists are another thing that calls for the researchers' attention. Frequency values are usually calculated as absolute occurrence of words, or normalized per million occurrence of a word to reduce the corpus size effect. These values might be biased because they do not take into account the Contextual Diversity (CD) of a word, which refers to the number of contexts (or passages, documents) a word appears in the source corpus that is used to compose the frequency norm. CD measures have been found to better account for the word frequency effects in Lexical Decision Tasks (LDT, Adelman et al., 2006). Compared to the absolute count of occurrence, the CD value is argued to be a fairer account because the more contexts in which a word occurs, the more likely it is to be encountered by language users and reinforced in their mental lexicon. However, whether the CD value is effective for predicting text readability is yet to be explored.

Last but not least, the way word frequency is used for readability assessment purpose also needs to be further investigated. Previous research mostly used simple average frequency count of words or percentage of words from the top frequency bands of the list to predict text readability. They were found to be successful to some extent. However, these methods are unable to capture the full picture of text readability from a word frequency perspective, because the averaging procedure is easily affected by extreme values and loses details. The method of counting the number of frequent words from the top bands of the frequency list neglects the contribution of less frequent words on the text's readability. It is precisely these words that are causing problems to the readers.

In light of the results from previous research, the present study tries to explore the relationship between word frequency of occurrence and text readability, seeking answers to the following questions:

1. How can word frequency, a lexical and local level characteristic of a text, be used to predict the text level characteristic of text readability?

2. Which frequency norms—the frequency lists and measures—are better predictors of text readability?

3. How can word frequency norms be better used to characterize text readability?

3.4 Methods and materials

Machine learning methods, which are a subfield of artificial intelligence that enables automatic construction of statistical models from data, were used in our experiments. Machine learning does not presuppose a statistical model to the data at hand. Rather, the models are automatically constructed by running some algorithms on the data. Two types of methods form the basics of machine learning: classification and clustering. The former solves problems related to assigning classes to new instances based on a set of features and the 'goldstandard' classes of the training instances. It is also called supervised learning because the classes of the instances used for training the statistical models are already known and used to 'supervise' the prediction of the classes of new instances. The latter, also known as unsupervised learning, refers to the process of grouping data instances without predefined classes. Both classification and clustering methods were used in the current research.

For classification, prediction accuracy is often used to evaluate the trained models. A 10-fold Cross Validation (CV) procedure can be applied to gain a better estimate of the model's performance. In 10-fold CV, the whole data set is evenly divided into 10 parts. Then, the same procedure is run 10 times by rotating the training and test sets to obtain an average accuracy from the 10 repetitions. For each repetition, 90% of the data are used for training, and the rest 10% for testing. In this study, besides 10-fold CV results, we also report the cross-corpus testing performance of the models trained with the whole training set and tested on a new set of data. This gives us a better impression of the models' generalizability across corpora. Because the training and test sets used different text leveling systems, Spearman's rank correlation (ρ) was used for evaluating the models' cross-corpus performance. Ten-fold CV ρs were also calculated for comparing within- and cross-corpus performances.

Detailed description of the corpora, the frequency lists, the features tested, and the experimental procedures are given later.

3.4.1 The WeeBit and Common Core corpora

The WeeBit corpus consists of reading passages from both the American educational magazine Weekly Reader and the BBC-Bitesize website. Texts in the corpus are labeled with one of the five reader levels that the original articles targeted at: Weekly

Reader Levels 2–4, BiteSize KS3, and BiteSize GCSE. The corpus consists of 789,926 words, with 616 texts in each of the five levels. Table 3.1 summarizes the profile of the corpus. Data from this corpus were used to train readability models.

Grade Level	Target Age Group	Number of Articles	Avg. Number of Words per Article
WR Level 2	7–8	616	152.63
WR Level 3	8–9	616	190.74
WR Level 4	9–10	616	294.91
BiteSize KS 3	11–14	616	243.56
BiteSize GCSE	14–16	616	400.51

Table 3.1: Details of the WeeBit corpus

The Common Core corpus consists of 168 texts given as sample texts appropriate for students from grades 2 to 12 in Appendix B of the English Language Arts Standards of the Common Core State Standards (CCSSO, 2010). It was used for testing the trained models in the current research.

3.4.2 The SUBTLEX frequency lists

In order to investigate the relationship between word frequency and text readability, a normative frequency list of the language is required. Because the frequency effects work on the readers' perception of words, the frequency list needs to be a faithful representation of their exposure to the vocabulary in the language. A careful comparison led us to the SUBTLEX frequency lists (Brysbaert and New, 2009; van Heuven et al., 2014), which stood out because of their recency and effectiveness in accounting for the latencies in naming and lexical decision tasks. A naming task requires participants to assign correct names to objects presented to them. In LDT, participants are asked if the stimuli are words or non-words. Both the two types of tasks have been used to test the participants' vocabulary knowledge of the language in psychology and psycholinguistic experiments. Furthermore, the SUBTLEX lists are based on spoken English corpora, which are a better reflection of language use in people's daily life than written ones because most people have more exposure to the spoken form of a language than its written form.

The SUBTLEXus (Brysbaert and New, 2009) list was constructed from a 51-million-word corpus consisting of subtitles from 8,388 American films and television series between the years 1900 and 2007. Brysbaert and New argued for their selection of the alternative source of language use from film and television subtitles by stating

that they 'usually involve people in social interactions' (p. 979), which happen more often than interactions with the written source to language users. As a result, their list was found to have stronger predictive power to vocabulary processing latencies in lexical decision and naming tasks than other frequency lists (e.g., the Kučera and Francis list).

The SUBTLEXuk (van Heuven et al., 2014) list is the British English version of SUBTLEXus. Its corpus was from the subtitles of nine British TV channels broadcast from January 2010 to December 2012, which consisted of 45,099 different broadcasts and 201.7 million words. Because the WeeBit corpus used in our experiments consisted of a mixture of British and American English texts, both the SUBTLEXus and SUBTLEXuk were used for comparison. The number of word forms included in the U.S. and U.K. lists are 74,286 and 160,022 words, respectively. Both lists included the raw occurrence count for each word form from the corresponding subtitle corpora, as well as contextual diversity measures and their corresponding normalized values. The U.K. list also provides frequency measures from its sub-corpora and the British National Corpus (BNC)[2] for comparison. Table 3.2 lists the SUBTLEX frequency measures that were used in this research and their descriptions.

3.4.3 Preprocess and feature calculation

The SUBTLEX frequency lists were first imported into a computer relational database. For each of the features listed in Table 3.2, a stratification procedure was applied, resulting in an extra set of measures signifying the relative position of each word in terms of its feature value relative to the range of that feature's values in the whole list. The scheme was to stratify the original frequency lists into varying numbers of bands based on each of the frequency measures.

Texts from both the WeeBit and the Common Core corpora were tokenized with the CoreNLP Tokenizer (Manning et al., 2014)—the same tokenizer used for composing the SUBTLEX frequency lists. A token is the original form of a word as it appears in a text, and a tokenizer is a computer program that automatically separates tokens in sentences. The SUBTLEX lists contain entries of words in their token forms from the subtitle corpora since Brysbaert and New (2009) found that the token forms were more informative than their corresponding lemma forms when they were used to account for LDT. Most word forms from the WeeBit corpus found matching entries from the frequency lists. On average, only 5.33% (SD = 3.76%) of the tokens in a text did not find a matching entry from the SUBTLEXus

[2] `http://www.natcorp.ox.ac.uk/`

List	Feature	Explanation
US	FREQCOUNT	number of times the word appears in the corpus
	CDCOUNT	number of films in which the word appears
	SUBTLWF	word frequency per million words
	LG10WF	$\log_{10}$(FREQCOUNT+1)
	SUBTLCD	percent of the films the word appears
	LG10CD	$\log_{10}$(CDCOUNT+1)
	ZIPF	$\log_{10}$[perMillion(FREQCOUNT+1)]+3
UK	FREQCOUNT	number of times the word appears in the corpus
	CBEEBIES_FREQ	number of times the word appears in the Cbeebies broadcasts sub-corpus
	CBBC_FREQ	number of times the word appears in the CBBC broadcasts sub-corpus
	BNC_FREQ	number of times the word appears in the British National Corpus
	LOGFREQ_ZIPF	Zipf value from the complete corpus
	LOGFREQCBEEBIES_-ZIPF	Zipf value from the Cbeebies sub-corpus
	LOGFREQCBBC_ZIPF	Zipf value from the CBBC sub-corpus
	LOGFREQBNC_ZIPF	Zipf value from the BNC corpus
	CD_COUNT	number of broadcasts in which the word appears
	CD_COUNT_CBEEBIES	number of broadcasts from the sub-corpus Cbeebies in which the word appears
	CD_COUNT_CBBC	number of broadcasts from the sub-corpus CBBC in which the word appears
	CD	percentage of broadcasts in which the word appears
	CD_CBEEBIES	percentage of broadcasts from the sub-corpus Cbeebies in which the word appears
	CD_CBBC	percentage of broadcasts from the sub-corpus CBBC in which the word appears

Table 3.2: Frequency measures from the SUBTLEX lists

frequency norm, while that from the U.K. counterpart was 3.58% (SD = 2.98%). Duplicate tokens are commonly found in a text. By removing the duplicate tokens, we obtained a list of word types used by a text. For each experiment conducted in this research, we constructed both type and token models for comparison purposes.

For each text, the following feature values were calculated and used as attributes to train the classification models:

1. Experiment 1: Mean and Standard Deviation (SD) of the frequency measures listed in Table 3.2

2. Experiment 2: Mean frequency or mean percentage of words from each frequency band of increasing fine-grainedness

3. Experiment 3: Branch means of the hierarchical cluster tree built with word frequency values

Multiple classification algorithms, including decision trees, support vector machines, and K-Nearest Neighbors (KNN) were tested for constructing the classification models. The KNN algorithm consistently outperformed the other algorithms in our experiments, so the results from this algorithm are reported in this study. The experiment setup consists of two components: a Java program that calculates the text features and an R (R Core Team, 2015) script that builds the models and test model generalizability. Full technical setup, source code, and experimental procedure are downloadable from the authors' Web page[3].

3.5 Experiment 1: Mean and standard deviation of frequencies as readability features

Experiment 1 aimed at comparing the effectiveness of the frequency measures provided by the SUBTLEX lists as text readability assessment features. In terms of method, the most conservative method of averaging the frequency values was adopted, but we also added the SD of the frequency values as a feature of a text. As a result, four models were constructed for each of the frequency measure listed in Table 3.2: token/type models with/without SD (± SD).

Table 3.3 shows the performance of the models trained with different frequency measures. In general, models trained with both the mean and SD features performed consistently better than those with only mean frequencies, be it type or token means.

[3] `http://xiaobin.ch/` or `http://purl.org/dm`

Within-corpus cross validation ρ and cross-corpus ρ showed that the +SD models also had better within- and cross-corpus validity. The reason for the +SD models' better performance was that they not only took into account an overall summary of the word frequencies, but also a summary of how widespread the frequencies were, resulting in more input information to the models. Although adding the SD feature seems an easy and obvious choice, to the best of our knowledge, no previous study has included SD in their predictive models.

Another obvious finding from the statistics is that the type models had uniformly better accuracy and validation performance than the token models. Take the LOGFREQCBEEBIES_ZIPF measure from the U.K. list as an example, the model trained with type frequencies, whose 10-fold CV accuracy estimate was 52%, performed 9% better than its token counterpart at 43%. Although both models' predictions for the validation and test sets were significantly correlated to the actual levels, the correlations of the type models (0.65 and 0.54 for within- and cross-corpus ρs) were stronger than those of the token models (0.52 and 0.45, respectively). The results of this experiment conformed with our hypothesis of a frequency effect on readability, which is further illustrated in Figure 3.2. The figure shows box plots of mean token or type Zipf values on reading levels of the texts from the Common Core corpus. The left panel, which plots the token means, does not show any pattern across levels. However, when plotted with type values (the right panel), a decreasing trend on mean Zipf values by increasing text difficulty can be clearly seen. The reason is that the more frequent words have a higher number of occurrence and give more weight to the mean than the less frequent ones, obscuring the frequency difference among texts of different readability levels. These results also echo Laufer and Ravenhorst-Kalovski (2010) finding that difficult texts have lower lexical coverage of frequent words.

As for the different frequency measures, the standardized measures (e.g., the logarithm and Zipf measures) had in general better performance than raw occurrence counts. For example, while the raw frequency count (FREQCOUNT) from the U.S. list worked comparatively well in terms of CV accuracy and Spearman's ρ, the trained models were barely transferable to the test corpus—cross-corpus ρs being insignificantly low, at -0.01 and 0.1 for token and type models, respectively. However, the Zipf value, which was standardized from raw frequency count, had both better accuracy and significant cross-corpus testing ρs.

The last finding of Experiment 1 was that the corpus from which the frequency list was constructed mattered when the frequency list was used to characterize text readability. Zipf values were provided by both the SUBTLEX U.S. and U.K. lists.

Table 3.3: Performance of models trained with different frequency measures.

List	Feature Set	10-fold CV Accuracy (%)		10-fold CV ρ		Cross-corpus ρ	
		Token(+/–SD)	Type(+/–SD)	Token(+/–SD)	Type(+/–SD)	Token(+/–SD)	Type(+/–SD)
US	FREQCOUNT	0.36/0.32	0.41/0.30	$0.31^{***}/0.22^{**}$	$0.38^{***}/0.12$	-0.01/0.03	0.1/0.12
	CDCOUNT	0.42/0.27	0.43/0.32	$0.45^{***}/0$	$0.46^{***}/0.19^{**}$	$0.3^{***}/0.06$	$0.32^{***}/0.31^{***}$
	SUBTLWF	0.36/0.32	0.41/0.30	$0.32^{***}/0.22^{***}$	$0.37^{***}/0.13$	-0.01/0.02	$0.07/0.17^{*}$
	LG10WF	0.40/0.24	0.44/0.32	$0.4^{***}/0$	$0.42^{***}/0.26^{***}$	0.34^{***}/-0.01	$0.35^{***}/0.33^{***}$
	SUBTLCD	0.42/0.26	0.42/0.32	0.46^{***}/-0.02	$0.44^{***}/0.19^{*}$	0.28^{***}/-0.27***	$0.33^{***}/0.22^{**}$
	LG10CD	0.37/0.24	0.42/0.33	$0.28^{***}/0.1$	$0.38^{***}/0.34^{***}$	$0.22^{**}/0.12$	$0.35^{***}/0.33^{***}$
	ZIPF_VALUE	0.40/0.24	0.43/0.32	0.4^{***}/-0.02	$0.42^{***}/0.26^{**}$	$0.34^{***}/0.03$	$0.35^{***}/0.33^{***}$
UK	FREQCOUNT	0.34/0.25	0.40/0.32	$0.25^{***}/0.11$	$0.32^{***}/0.21^{***}$	0.15/-0.05	$0.18^{*}/0.36^{***}$
	CBEEBIES_FREQ	0.35/0.25	0.37/0.32	$0.24^{***}/0.07$	$0.25^{***}/0.2^{**}$	0.09/0.06	$0.21^{**}/0.26^{***}$
	CBBC_FREQ	0.36/0.26	0.39/0.31	$0.25^{***}/0.1$	$0.28^{***}/0.14$	$0.18^{*}/0.05$	$0.15/0.27^{***}$
	BNC_FREQ	0.31/0.26	0.40/0.34	$0.24^{***}/0.12$	$0.33^{***}/0.27^{***}$	$0.18^{*}/0.04$	$0.12/0.27^{***}$
	LOGFREQ_ZIPF	0.34/0.23	0.40/0.31	$0.24^{***}/0.05$	$0.39^{***}/0.33^{***}$	0.25^{**}/-0.03	$0.39^{***}/0.39^{***}$
	LOGFREQCBEEBIES_-ZIPF	0.43/0.30	0.52/0.41	$0.52^{***}/0.25^{***}$	$0.65^{***}/0.57^{***}$	$0.45^{***}/0.21^{**}$	$0.54^{***}/0.47^{***}$
	LOGFREQCBBC_ZIPF	0.37/0.24	0.44/0.35	$0.33^{***}/0.07$	$0.5^{***}/0.39^{***}$	$0.41^{***}/0.09$	$0.42^{***}/0.45^{***}$
	LOGFREQBNC_ZIPF	0.29/0.24	0.33/0.29	0.15/0.07	$0.23^{**}/0.17$	$-0.17^{*}/0.02$	$0.29^{***}/0.29^{***}$
	CD_COUNT	0.36/0.26	0.38/0.31	$0.27^{***}/0.04$	$0.3^{***}/0.22^{*}$	0.05/-0.07	$0.29^{***}/0.34^{***}$
	CD_COUNT_CBEEBIES	0.37/0.32	0.37/0.30	$0.33^{***}/0.1$	$0.27^{***}/0.13$	0.1/-0.07	$0.37^{***}/0.14$
	CD_COUNT_CBBC	0.41/0.29	0.41/0.30	$0.43^{***}/0.07$	$0.42^{***}/0.17^{*}$	0.17^{*}/-0.13	$0.36^{***}/0.28^{***}$
	CD	0.35/0.26	0.38/0.31	$0.28^{***}/0.04$	$0.3^{***}/0.22^{**}$	0.02/-0.09	$0.25^{***}/0.33^{***}$
	CD_CBEEBIES	0.37/0.32	0.37/0.30	$0.33^{***}/0.12$	$0.27^{***}/0.12$	0.12/-0.07	$0.37^{***}/0.12$
	CD_CBBC	0.42/0.29	0.41/0.30	$0.43^{***}/0.08$	$0.41^{***}/0.17^{*}$	0.17^{*}/-0.12	$0.37^{***}/0.3^{***}$

Token: performance of models trained with token data; Type: performance of models trained with type data; +/–SD: with/without standard deviation as a feature; ***: $p \leq 0.001$; **: $p \leq 0.01$; *: $p \leq 0.05$

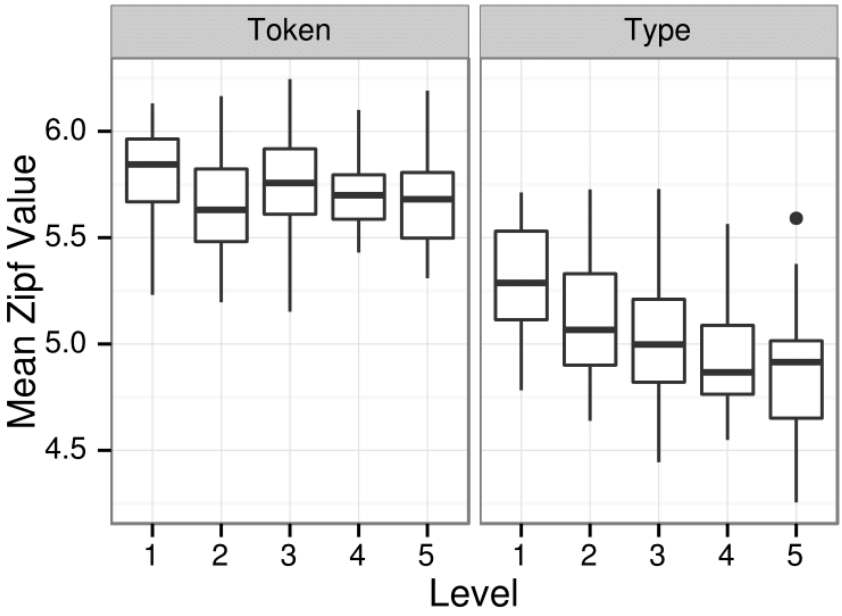

Figure 3.2: Mean type/token Zipf value by reading level

The U.K. list also included Zipf values calculated from the various sub-corpora. Models trained with Zipf values from different sources had varying performance. The CBEEBIES_ZIPF +SD type model not only had the best within-corpus accuracy and ρ (0.52 and 0.65, respectively), but also was most generalizable to the test corpus, with a cross-corpus ρ of 0.54 ($p < .001$). As reviewed earlier, a frequency list that reflects the readers' actual usage and experience of the language would be an effective reference for assessment of text readability. The optimal performance of the CBEEBIES_ZIPF measure shows that the measure is not only a faithful representation of the actual language use by the subcorpus' target group, but also reflects, at least to some degree, language development because it is calculated from TV programs aiming at school children who are emergent readers.

In all, the results of this experiment suggested that in order to guarantee optimal accuracy and generalizability, the classification model needs to be constructed with standardized word type frequency norms obtained from a corpus that reflects actual language use and development.

3.6 Experiment 2: Mean frequencies of words from language frequency bands with increasing fine-grainedness

Another way to characterize text readability from the word frequency perspective is to divide the frequency range into differing number of bands and calculate the mean frequencies of words from each of these bands. For comparison purpose, the percentage of words from each band was also calculated and used to train the readability models. The hypothesis of this experiment is that the more words of a text are from the less frequent bands, the higher the perception demand for these words, hence higher textual difficulty and lower readability. An analogy of the experiment is to measure readability with the 'ruler' of general language use—the frequency lists. One problem research needs to solve is how to construct such a ruler: how fine-grained the calibration points need to be. That is to say, in this experiment, we aimed at deciding how many bands the frequency list needed to be divided into for the trained model to achieve the highest predictive accuracy and cross-corpus generalizability.

Results from Experiment 1 revealed the optimal performance of the type models trained with the Zipf value and SUBTLCD measures from the SUBTLEXus list. The two measures represent two ways in which frequency can be measured: frequency of occurrence and contextual distribution. As a result, Experiment 2 was conducted on the basis of these measures. For comparison purposes, the same measures from the SUBTLEXuk list were also used for model construction.

We started by dividing the frequency lists into two halves and gradually increased the fine-grainedness of the calibration. For example, for the Zipf measure of word frequency from the U.S. list, the maximum value of a word is 7.62 and the minimum 1.59. When the list was divided into two halves based on the Zipf values of word entries, words with Zipf values between 1.59 and 3.02 were in the lower band and those with Zipf values higher than 3.02 were in the upper band. When it was stratified into three bands, the band ranges became 1.59–3.6, 3.6–5.61, and 5.61–7.62, resulting in a finer-grained calibration. In this study, we experimented with up to 100 bands for the selected frequency measures. For each text in the training and testing corpora, mean frequencies of the words used by the text in each frequency band were calculated and stored as attributes of the text. A separate experiment used percentage of words from each frequency band as text features. Classifiers were then trained on the band averages or percentages statistics and their performance evaluated.

Results of the experiment are shown in Tables 3.4–3.7, which suggest that the performance of models trained with mean frequencies of words from each frequency band and those trained with percentage of words from each band was not significantly different. Similar patterns were found from both the two methods. First, with the increase of band numbers, the within-corpus ρs keep increasing, while the cross-corpus ρs remain stable. Second, the CD measures did not perform as well as the ZIPF measures in either the within-corpus 10-fold CV evaluation or the cross-corpus testing. In terms of methods, the mean frequency method is in favor of the type models, while the percentage method performed better with token models. These findings echoed those from Experiment 1 that type frequencies are better estimation of text readability than token frequencies.

Table 3.6 shows that despite their consistent within-corpus performance improvement with the increasing number of stratification bands, models trained on type values of the U.K. list had little generalizability. The 'NA' s in Table 3.6 mean that the trained models failed to calculate Spearman's ρ because the prediction made by them was homogeneous, violating the nonzero SD requirement of correlation coefficient calculations. This also suggests that the trained models were incapable of distinguishing the reading levels of the testing texts. Similar findings were obtained with the percentage method (Table 3.7). The cross-corpus performance of the U.S. models were also better than that of the U.K. models.

Figure 3.3 plots a comparison of the performance of different models constructed with the SUBTLEXus measures. Models built with Zipf values had better training and testing performance than those with the CD values. The figure also shows that finer-grained frequency bands did not improve the generalizability of the trained models beyond 20 bands, which suggests that cutting the frequency list into finer-grained bands is not necessary when it comes to using it for readability assessment.

The results of Experiment 2 show the effectiveness of using frequency lists as 'rulers' of language use to measure readability. Both the SUBTLEX U.S. and U.K. lists were effective in measuring the training corpus with increased fine-grainedess of calibration. However, stratifying the frequency list into more than 20 bands did not improve model performance. The U.S. list had better performance when the trained models were carried over to a test corpus. Consequently, depending on the purpose of application and which frequency measure to use, one still needs to consider how fine-grained the frequency list needs to be stratified.

Table 3.4: Performance of models trained on mean frequency of words from stratified frequency bands of two SUBTLEXus measures

No. of Bands	ZIPF (Token/Type)			CD (Token/Type)		
	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ
2	0.29/0.33	0.17/0.26	$0.2^{*}/0.37^{***}$	0.35/0.33	$0.16^{*}/0.16^{*}$	0.11/0.09
3	0.30/0.33	$0.19^{**}/0.24^{**}$	$0.19^{*}/0.4^{***}$	0.37/0.36	$0.29^{***}/0.25^{***}$	$0.17^{*}/0.21^{**}$
4	0.30/0.34	$0.21^{***}/0.27^{***}$	$0.3^{***}/0.35^{***}$	0.37/0.37	$0.23^{**}/0.24^{**}$	0.09/0.1
5	0.30/0.34	$0.2^{**}/0.25^{**}$	$0.26^{***}/0.31^{***}$	0.44/0.44	$0.47^{***}/0.49^{***}$	$0.2^{*}/0.35^{***}$
6	0.31/0.34	$0.23^{**}/0.28^{**}$	$0.3^{***}/0.35^{***}$	0.45/0.46	$0.52^{***}/0.5^{***}$	$0.17^{*}/0.29^{***}$
7	0.30/0.33	$0.23^{***}/0.26^{***}$	$0.32^{***}/0.34^{***}$	0.46/0.46	$0.58^{***}/0.54^{***}$	$0.23^{**}/0.35^{***}$
8	0.31/0.35	$0.24^{***}/0.29^{***}$	$0.33^{***}/0.38^{***}$	0.48/0.46	$0.6^{***}/0.53^{***}$	$0.22^{**}/0.19^{*}$
9	0.32/0.34	$0.25^{***}/0.29^{***}$	$0.33^{***}/0.37^{***}$	0.50/0.48	$0.63^{***}/0.56^{***}$	$0.37^{***}/0.38^{***}$
10	0.32/0.34	$0.25^{***}/0.29^{***}$	$0.32^{***}/0.36^{***}$	0.51/0.47	$0.63^{***}/0.54^{***}$	$0.37^{***}/0.15$
20	0.53/0.55	$0.8^{***}/0.81^{***}$	0.07/0	0.47/0.45	$0.47^{***}/0.44^{***}$	$0.22^{**}/0.19^{*}$
30	0.61/0.61	$0.82^{***}/0.83^{***}$	$0.42^{***}/0.3^{***}$	0.46/0.46	$0.55^{***}/0.53^{***}$	$0.17^{*}/0.16^{*}$
40	0.61/0.60	$0.82^{***}/0.83^{***}$	$0.37^{***}/0.36^{***}$	0.47/0.47	$0.49^{***}/0.5^{***}$	$0.17^{*}/0.15$
50	0.62/0.63	$0.84^{***}/0.84^{***}$	$0.34^{***}/0.34^{***}$	0.47/0.47	$0.49^{***}/0.48^{***}$	$0.17^{*}/0.19^{*}$
60	0.62/0.62	$0.83^{***}/0.83^{***}$	$0.39^{***}/0.38^{***}$	0.49/0.49	$0.58^{***}/0.56^{***}$	$0.13/0.2^{**}$
70	0.61/0.61	$0.82^{***}/0.82^{***}$	$0.4^{***}/0.41^{***}$	0.49/0.49	$0.56^{***}/0.56^{***}$	$0.16^{*}/0.19^{*}$
80	0.62/0.62	$0.83^{***}/0.82^{***}$	$0.39^{***}/0.37^{***}$	0.53/0.53	$0.69^{***}/0.69^{***}$	$0.2^{**}/0.18^{*}$
90	0.62/0.62	$0.81^{***}/0.82^{***}$	$0.29^{***}/0.37^{***}$	0.54/0.54	$0.67^{***}/0.67^{***}$	$0.13/0.15^{*}$
100	0.62/0.62	$0.82^{***}/0.83^{***}$	$0.4^{***}/0.37^{***}$	0.56/0.57	$0.7^{***}/0.67^{***}$	$0.23^{**}/0.25^{**}$

***: $p \leq 0.001$; **: $p \leq 0.01$; *: $p \leq 0.05$

Table 3.5: Performance of models trained on percentage of words from stratified frequency bands of two SUBTLEXus measures

No. of Bands	ZIPF (Token/Type)			CD (Token/Type)		
	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ
2	0.26/0.33	$0.26^{***}/0.25^{***}$	$0.38^{***}/0.4^{***}$	0.29/0.33	$0.22^{*}/0.13^{*}$	0.07/0.14
3	0.26/0.33	$0.26^{***}/0.24^{***}$	$0.37^{***}/0.4^{***}$	0.32/0.37	$0.39^{***}/0.25^{***}$	$0.24^{**}/0.21^{**}$
4	0.28/0.34	$0.29^{***}/0.28^{***}$	$0.35^{***}/0.32^{***}$	0.36/0.37	$0.42^{***}/0.23^{***}$	$0.24^{**}/0.09$
5	0.28/0.34	$0.29^{***}/0.27^{***}$	$0.36^{***}/0.35^{***}$	0.37/0.44	$0.45^{***}/0.51^{***}$	$0.22^{**}/0.27^{***}$
6	0.29/0.34	$0.3^{***}/0.29^{***}$	$0.37^{***}/0.35^{***}$	0.38/0.46	$0.46^{***}/0.54^{***}$	$0.2^{*}/0.21^{**}$
7	0.30/0.33	$0.31^{***}/0.27^{***}$	$0.37^{***}/0.33^{***}$	0.40/0.45	$0.51^{***}/0.53^{***}$	$0.28^{***}/0.28^{***}$
8	0.30/0.34	$0.3^{***}/0.28^{***}$	$0.32^{***}/0.37^{***}$	0.43/0.46	$0.52^{***}/0.54^{***}$	$0.24^{**}/0.24^{**}$
9	0.31/0.34	$0.31^{***}/0.29^{***}$	$0.37^{***}/0.36^{***}$	0.42/0.48	$0.52^{***}/0.58^{***}$	$0.25^{**}/0.3^{***}$
10	0.31/0.34	$0.31^{***}/0.3^{***}$	$0.38^{***}/0.35^{***}$	0.43/0.48	$0.52^{***}/0.56^{***}$	$0.27^{***}/0.19^{*}$
20	0.54/0.55	$0.77^{***}/0.81^{***}$	$0.29^{***}/0$	0.45/0.45	$0.49^{***}/0.43^{***}$	$0.16^{*}/0.19^{*}$
30	0.53/0.61	$0.73^{***}/0.83^{***}$	$0.26^{***}/0.43^{***}$	0.49/0.46	$0.53^{***}/0.52^{***}$	$0.3^{***}/0.12$
40	0.52/0.60	$0.71^{***}/0.83^{***}$	$0.28^{***}/0.38^{***}$	0.48/0.47	$0.52^{***}/0.49^{***}$	$0.2^{*}/0.14$
50	0.52/0.63	$0.69^{***}/0.84^{***}$	$0.21^{**}/0.36^{***}$	0.50/0.47	$0.57^{***}/0.48^{***}$	$0.19^{*}/0.19^{*}$
60	0.52/0.63	$0.68^{***}/0.83^{***}$	$0.32^{***}/0.38^{***}$	0.52/0.49	$0.6^{***}/0.58^{***}$	$0.22^{**}/0.22^{**}$
70	0.55/0.62	$0.71^{***}/0.82^{***}$	$0.07/0.39^{***}$	0.53/0.49	$0.59^{***}/0.56^{***}$	-0.06/0.11
80	0.57/0.62	$0.69^{***}/0.82^{***}$	$0.15^{*}/0.37^{***}$	0.54/0.53	$0.65^{***}/0.7^{***}$	$0.12/0.17^{*}$
90	0.57/0.62	$0.7^{***}/0.82^{***}$	$0.2^{**}/0.35^{***}$	0.54/0.54	$0.67^{***}/0.67^{***}$	$0.05/0.18^{*}$
100	0.58/0.62	$0.72^{***}/0.83^{***}$	$0.06/0.36^{***}$	0.55/0.57	$0.66^{***}/0.68^{***}$	$0.24^{**}/0.24^{**}$

***: $p \leq 0.001$; **: $p \leq 0.01$; *: $p \leq 0.05$

Table 3.6: Performance of models trained on mean frequency of words from stratified frequency bands of two SUBTLEXuk measures

No. of Bands	ZIPF (Token/Type)			CD (Token/Type)		
	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ
2	0.25/0.33	0.04/0.28	0.04/NA	0.32/0.33	0.16^{*}/0.26^{*}	0.17^{*}/NA
3	0.26/0.34	0.06/0.27	0.18^{*}/NA	0.34/0.35	0.2^{**}/0.24^{**}	0.04/NA
4	0.25/0.35	0.05/0.3	-0.01/NA	0.39/0.39	0.39^{***}/0.36^{***}	0.1/NA
5	0.45/0.52	0.68^{***}/0.74^{***}	0.33^{***}/NA	0.39/0.40	0.37^{***}/0.38^{***}	0.07/NA
6	0.45/0.52	0.69^{***}/0.74^{***}	0.29^{***}/-0.05	0.38/0.38	0.33^{***}/0.29^{***}	0.06/NA
7	0.45/0.52	0.69^{***}/0.74^{***}	0.3^{***}/-0.05	0.41/0.40	0.42^{***}/0.39^{***}	0.15/NA
8	0.46/0.52	0.68^{***}/0.74^{***}	0.3^{***}/-0.05	0.42/0.42	0.41^{***}/0.37^{***}	0.18^{*}/NA
9	0.46/0.53	0.69^{***}/0.73^{***}	0.35^{***}/NA	0.43/0.42	0.44^{***}/0.4^{***}	0.2^{*}/NA
10	0.47/0.53	0.69^{***}/0.74^{***}	0.38^{***}/NA	0.43/0.42	0.4^{***}/0.38^{***}	0.22^{**}/NA
20	0.49/0.51	0.74^{***}/0.76^{***}	-0.14/NA	0.52/0.50	0.64^{***}/0.64^{***}	0.12/0.05
30	0.56/0.55	0.77^{***}/0.78^{***}	0.03/NA	0.51/0.51	0.59^{***}/0.59^{***}	0.2^{**}/-0.08
40	0.60/0.60	0.8^{***}/0.81^{***}	0.38^{***}/0.15^{*}	0.52/0.52	0.64^{***}/0.64^{***}	0.23^{**}/-0.3^{***}
50	0.61/0.61	0.8^{***}/0.81^{***}	0.33^{***}/0.17^{*}	0.52/0.52	0.6^{***}/0.61^{***}	0.22^{**}/0.21^{**}
60	0.61/0.62	0.79^{***}/0.79^{***}	0.36^{***}/0.1	0.55/0.56	0.72^{***}/0.73^{***}	0.25^{**}/-0.01
70	0.64/0.64	0.8^{***}/0.81^{***}	0.28^{***}/0.3^{***}	0.55/0.55	0.7^{***}/0.71^{***}	0.22^{**}/-0.13
80	0.62/0.63	0.8^{***}/0.8^{***}	0.31^{***}/0.1	0.55/0.56	0.75^{***}/0.75^{***}	0.23^{**}/0.17^{*}
90	0.62/0.62	0.8^{***}/0.8^{***}	0.32^{***}/-0.32^{***}	0.57/0.57	0.75^{***}/0.75^{***}	0.27^{***}/NA
100	0.59/0.59	0.78^{***}/0.76^{***}	0.22^{**}/-0.1	0.58/0.58	0.74^{***}/0.74^{***}	0.24^{**}/-0.02

NA: unable to calculate ρ because sd equals 0, trained model not generalizable; ***: $p \leq 0.001$; **: $p \leq 0.01$; *: $p \leq 0.05$

Table 3.7: Performance of models trained on SUBTLEXuk features with percentage of words from stratified frequency bands

No. of Bands	ZIPF (Token/Type)			CD (Token/Type)		
	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ
2	0.22/0.34	0.04/0.28	0.17*/NA	0.24/0.33	0.04/0.24	-0.01/NA
3	0.24/0.35	0.06/0.28	-0.16*/NA	0.28/0.35	0.23***/0.24***	0.17*/NA
4	0.24/0.35	0.05/0.29	-0.13/NA	0.37/0.39	0.42***/0.37***	0.09/NA
5	0.45/0.52	0.72***/0.73***	0.27***/NA	0.38/0.39	0.45***/0.35***	0.06/NA
6	0.46/0.52	0.72***/0.74***	0.25***/-0.07	0.40/0.38	0.43***/0.29***	0.09/NA
7	0.46/0.52	0.71***/0.74***	0.24**/-0.05	0.41/0.41	0.49***/0.4***	0.06/NA
8	0.47/0.52	0.72***/0.74***	0.26***/-0.05	0.42/0.42	0.49***/0.4***	0.02/NA
9	0.47/0.53	0.72***/0.73***	0.23**/NA	0.44/0.43	0.5***/0.4***	0/NA
10	0.47/0.52	0.71***/0.74***	0.28***/NA	0.44/0.42	0.49***/0.38***	-0.02/NA
20	0.50/0.51	0.73***/0.76***	-0.05/0.08	0.47/0.51	0.49***/0.61***	0/0.06
30	0.54/0.55	0.71***/0.78***	0.18*/NA	0.49/0.50	0.53***/0.59***	0.11/0.11
40	0.51/0.60	0.66***/0.81***	0.27***/0.15	0.51/0.52	0.57***/0.64***	0.04/-0.24**
50	0.47/0.61	0.6***/0.8***	0.36***/0.14	0.52/0.52	0.57***/0.61***	0.16*/0.22**
60	0.49/0.61	0.57***/0.78***	0.15*/0.16*	0.52/0.55	0.59***/0.72***	0.18*/-0.12
70	0.52/0.63	0.62***/0.81***	0.19*/0.31***	0.53/0.55	0.6***/0.71***	0.13/-0.01
80	0.53/0.63	0.62***/0.8***	0.17*/0.12	0.54/0.56	0.62***/0.74***	0.12/-0.07
90	0.54/0.62	0.63***/0.8***	0.31***/-0.33***	0.55/0.56	0.64***/0.74***	0.22**/-0.11
100	0.56/0.59	0.67***/0.76***	0.16*/-0.18*	0.57/0.58	0.6***/0.73***	0.21**/-0.22**

NA: unable to calculate ρ because sd equals 0, trained model not generalizable; ***: $p \leq 0.001$; **: $p \leq 0.01$; *: $p \leq 0.05$

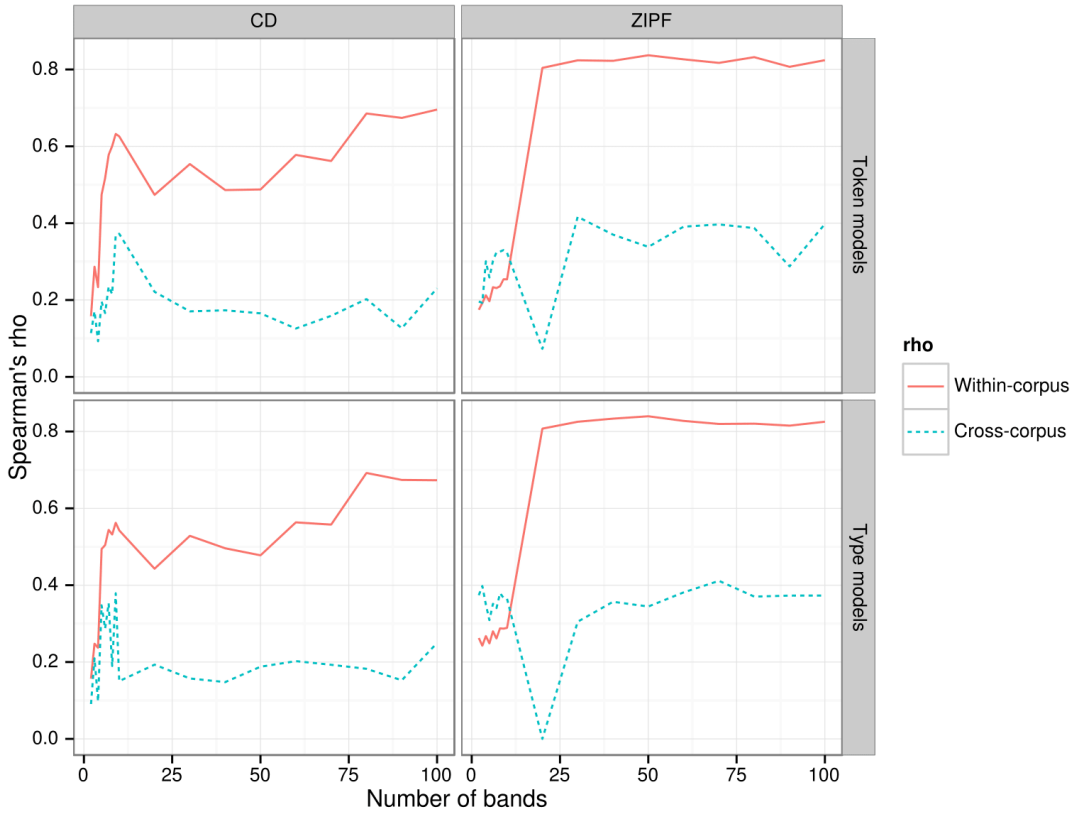

Figure 3.3: Performance of models trained on stratification schemes with measures from SUBTLEXus

3.7 Experiment 3: Predicting readability with frequency cluster means

In the cases of Experiments 1 and 2, text readability was characterized from an 'external' perspective, namely the frequency norms. Readability prediction is also approachable from an 'internal' perspective, in which the frequency distribution of the words used in a text is considered by themselves. With this approach, the frequency lists are not divided into bands based on certain frequency values. Rather, the frequency values of all words used in a text are obtained and clustered, resulting in the words with similar frequency values being grouped together. Clustering is an unsupervised machine learning technique that groups objects with similar characteristics together. It does not presuppose any classification of the objects but group them based on how close they are with each other in terms of the interested measures. Depending on the application, the number of clusters in which the objects are grouped is configurable.

In this experiment, Zipf values from the SUBTLEX lists were obtained for each word in a text. A Zipf value hierarchical clustering tree was created for each text with the hierarchical clustering algorithm `hclus()` provided by R. The clustering tree was then cut into different number of branches. The number of branches a tree was cut into represented the number of clusters the words in a text were grouped based on their Zipf values. Because the easiest level of the WeeBit corpus had the lowest average length of 152 words per text, the number of clusters tested was limited to a maximum of 100 clusters to avoid having too many clusters with single words. For each cluster, the average Zipf value of the words in that cluster was stored as a feature of the text. The feature set thus created from the cluster tree was used for supervised classifier training. We experimented on the cutting schemes, trying to find the optimal number of branches the cluster tree should be cut into.

Table 3.8 shows part of the performance results of the classifiers trained with different numbers of clusters. Due to space limit, Table 3.8 only shows performance of models with number of clusters divisible by 10 when the cluster tree is cut into more than 10 clusters. Figure 3.4 compares the performance of models with measures from the two different frequency lists.

The results show that the type and token models did not perform significantly different in terms of accuracy estimates, within- or cross-corpus ρs. Nor did we find significant differences between the performance of models trained on measures from different lists. All models showed an improved performance with the increase of cluster numbers. However, despite the continuous increase of within-corpus accuracy

estimates, the correlation between the predicted and actual reading levels did not increase consistently. For both the token and type models, the curves peaked at around 70 clusters. Besides, the Zipf measure from the U.S. list performed marginally better than its counterpart from the U.K. list in terms of testing results and cross-corpus ρ.

What is interesting about these results is that the trained classifiers are mostly generalizable to the test corpus, which is another confirmation of the existence of a frequency effect on readability. Words chosen by texts of different reading levels are characteristic of their difficulty. The better performance of the models in this study also suggests the superiority of the method used in this experiment than the one in Experiment 2. In Experiment 2, readability was measured with the 'ruler' of the frequency norms. However, in the third experiment, the problem was approached from the perspective of the documents themselves—grouping word usages by way of word frequencies. The reason for the better performance of the latter method coincides with findings from Experiment 1, which revealed a trend of less-frequent words being used more often in texts of higher reading levels. Hence, the cluster averages were able to capture the characteristics of word usage in texts of varying difficulty levels.

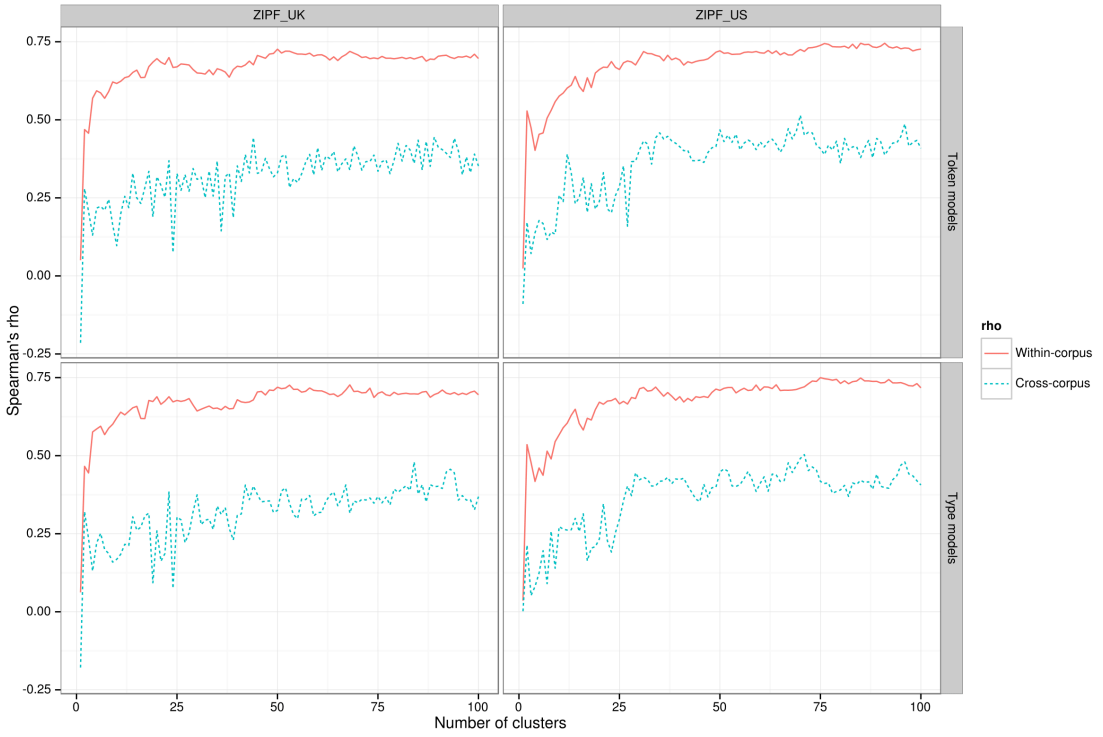

Figure 3.4: Performance of models trained on cluster schemes with Zipf measures from the SUBTLEX frequency lists

Table 3.8: Performance of type and token models trained on cluster mean Zipf values

No. of Clusters	SUBTLEXus ZIPF (Token/Type)			SUBTLEXuk ZIPF (Token/Type)		
	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ	CV(10-fold) Acc. (%)	Within-corpus ρ	Cross-corpus ρ
1	0.25/0.24	0.02/0.04	-0.09/0	0.24/0.24	0.05/0.06	-0.21^{**}/-0.18^{*}
2	0.39/0.39	0.53^{***}/0.54^{***}	0.17^{*}/0.21^{**}	0.37/0.37	0.47^{***}/0.47^{***}	0.28^{***}/0.32^{***}
3	0.35/0.35	0.47^{***}/0.48^{***}	0.07/0.05	0.37/0.37	0.46^{***}/0.44^{***}	0.21^{**}/0.23^{**}
4	0.36/0.36	0.4^{***}/0.42^{***}	0.14/0.08	0.42/0.42	0.57^{***}/0.58^{***}	0.13/0.13
5	0.39/0.39	0.45^{***}/0.46^{***}	0.18^{*}/0.12	0.44/0.44	0.59^{***}/0.59^{***}	0.22^{**}/0.22^{**}
6	0.39/0.39	0.46^{***}/0.44^{***}	0.17^{*}/0.2^{*}	0.44/0.44	0.59^{***}/0.59^{***}	0.22^{**}/0.25^{**}
7	0.42/0.42	0.51^{***}/0.51^{***}	0.12/0.09	0.44/0.44	0.57^{***}/0.57^{***}	0.21^{**}/0.2^{**}
8	0.44/0.43	0.53^{***}/0.49^{***}	0.14/0.26^{***}	0.45/0.45	0.59^{***}/0.59^{***}	0.25^{**}/0.19^{*}
9	0.44/0.44	0.56^{***}/0.55^{***}	0.13/0.14	0.45/0.46	0.62^{***}/0.6^{***}	0.16^{*}/0.16^{*}
10	0.45/0.45	0.58^{***}/0.57^{***}	0.26^{***}/0.27^{***}	0.45/0.45	0.62^{***}/0.62^{***}	0.1/0.17^{*}
20	0.46/0.48	0.66^{***}/0.67^{***}	0.24^{**}/0.23^{**}	0.49/0.49	0.7^{***}/0.69^{***}	0.32^{***}/0.26^{***}
30	0.48/0.48	0.7^{***}/0.71^{***}	0.4^{***}/0.42^{***}	0.49/0.49	0.65^{***}/0.64^{***}	0.31^{***}/0.37^{***}
40	0.47/0.47	0.69^{***}/0.69^{***}	0.41^{***}/0.42^{***}	0.49/0.49	0.67^{***}/0.68^{***}	0.35^{***}/0.31^{***}
50	0.50/0.51	0.72^{***}/0.71^{***}	0.47^{***}/0.45^{***}	0.53/0.53	0.73^{***}/0.72^{***}	0.33^{***}/0.32^{***}
60	0.50/0.50	0.71^{***}/0.71^{***}	0.43^{***}/0.41^{***}	0.53/0.53	0.71^{***}/0.71^{***}	0.41^{***}/0.32^{***}
70	0.52/0.51	0.73^{***}/0.72^{***}	0.51^{***}/0.49^{***}	0.53/0.53	0.71^{***}/0.71^{***}	0.38^{***}/0.35^{***}
80	0.53/0.54	0.73^{***}/0.73^{***}	0.36^{***}/0.39^{***}	0.53/0.53	0.7^{***}/0.7^{***}	0.42^{***}/0.38^{***}
90	0.54/0.53	0.74^{***}/0.73^{***}	0.43^{***}/0.4^{***}	0.54/0.54	0.7^{***}/0.7^{***}	0.41^{***}/0.4^{***}
100	0.54/0.54	0.73^{***}/0.72^{***}	0.41^{***}/0.41^{***}	0.53/0.53	0.7^{***}/0.69^{***}	0.35^{***}/0.37^{***}

***: $p \leq 0.001$; **: $p \leq 0.01$; *: $p \leq 0.05$

3.8 Summary

The purpose of the series of experiments conducted in this research was to confirm the possibility of characterizing text readability with the lexical measure of word frequency. In order to explore the connection between word frequency and readability assessment, NLP and ML technologies were employed. The experiments were carried out with two corpora with texts labeled with reading levels and two frequency norms that provided not only raw occurrence counts of words but also contextual diversity measures and their corresponding normalized values. A comparison of the effects of the different measures on readability assessment was carried out to determine which measures are better predictors of text difficulty. Besides the simple methods of calculating the frequency means and SD of all the words in a text, two other ways to characterize text readability were also tested. In the first approach, the frequency lists were divided into gradually more fine-grained levels of vocabulary bands. Text readability was then characterized by calculating the mean frequency values or the percentage of words from each band. The second method clusters the words in a text based on their frequencies in the frequency norms and then characterizes the text readability in terms of the cluster means.

The results of the experiments revealed that the choice of frequency lists, the frequency measures, and the way they were used to characterize text readability resulted in different performance of the readability classifiers. The fact that different frequency measures from different frequency lists had different model performance suggested that they differ in how faithfully they represent language experience. The normalized measures, such as the Zipf measure from the SUBTLEX frequency lists, seem to provide a more accurate estimate of the cognitive load involved in vocabulary perception and retrieval. The higher cognitive demand of successfully reading more difficult texts is traceable to the higher cognitive demand in understanding the words used in the text. As a result, the more a frequency list is capable of predicting the ease of vocabulary retrieval from the mental lexicon, the more useful it is for predicting text readability. The sub-corpus CBEEBIES of the SUBTLEXuk corpus is composed of children TV program subtitles, which are representative of language exposure for children developing their language proficiencies. The frequency measures from such a list are more likely to reflect the cognitive load from vocabulary retrieval and perception of the readers; hence, the readability models built on them had the highest predictive power (see results of Experiment 1).

As for the question of how to better characterize readability with frequency measures, the studies in this research showed that although the stratification method had improved within-corpus accuracy, its generalizability was limited. The cluster-

ing experiments showed better generalizability of the trained models, but they were also computationally more expensive than the other methods. By fine-tuning the number of bands into which the frequency lists are cut or the number of clusters to group the words in a text, it is possible to obtain more accurate and generalizable models than using the simple average used by most approaches. The clustering scheme in Experiment 3 yields the best models. They are the least sensitive to the frequency lists and the frequency measures. However, they are also the most difficult to calculate. As a result, while considering using word frequency to assess text readability, one needs to take into account the various aspects involved, namely, the frequency list, the frequency measures and the method used to aggregate the lexical information at the text level.

The results of the series of experiments is promising for the future of readability assessment based on textual measures. A single measure of word frequency was capable of achieving an estimation accuracy of more than 60%, which is comparable to other experiments using a combination of a number of semantic and syntactic measures (see Nelson et al., 2012, for evaluation of established readability assessment systems). However, we also acknowledge that for real-life application of a readability classifier trained on textual features more research still needs to be done. The open CTAP platform (Chen and Meurers, 2016b, see also Chapter 2) readily supports exploring a broad range of readability features. More focus should be placed on more comprehensively characterizing the language in a text in terms of its morphology, syntax, and semantics, and the cognitive demands of the reading process. The combination of these features whose roles in readability assessment need to be more fully understood will enable a more comprehensive characterization of text readability.

Chapter 4

Complexity and Accuracy

Chapter highlights

What is already known about this topic:

- Complexity, Accuracy, and Fluency (CAF) have become the central foci of language acquisition research and have been systematically used to evaluate language proficiency and development.
- In terms of language development, Skehan's Trade-Off Hypothesis (TOH) and Robinson's Cognition Hypothesis (CH) have different predictions on the developmental interrelationship between the CAF constructs. Both hypotheses have found supports from empirical studies, but have also been rejected by others, making it difficult to draw conclusions on the issue.
- Automatic tools such as the CTAP (see Chapter 2) have made it possible to investigate the multiple dimensions of complexity development in great detail.

What this study adds:

- The developmental inter-relationship between complexity and accuracy was accounted for with comprehensive sets of complexity and accuracy measures, making it possible to observe the development of the both constructs from fine-grained perspectives.
- We also compared the developmental inter-relationship between accuracy and complexity in both L1 and L2.
- Unlike previous studies which were mainly intervention studies that varied task factors, the current study uses longitudinal data from natural instructional

settings, making the findings on the developmental inter-relationship between the CAF constructs more convincing.

Implications for theory, policy, or practice:

- For theory:
  - Results from the study supported the simultaneous development of CAF constructs as postulated by the CH, but reject the TOH.
  - Development was observable for both accuracy and complexity in both L1 and L2 during the course of a typical four-month semester. The L1 group developed mainly at the higher textual levels of cohesion, rhetorics, and explicitness of language expression, while the L2 group showed development mainly at the lower linguistic levels of lexis and syntax.
- For practice: It is not necessary for language instructors to prioritize any one aspect—complexity and accuracy can develop simultaneously.

Abstract

While measures of CAF have been successfully used to study language development, perspectives differ on the relationship between complexity and accuracy. On the applied side, questions also arise regarding the impact of writing instruction on the development of the L2 as compared to the L1. To address these conceptual and applied issues, this study investigates the developmental patterns in a longitudinal corpus of adult student writings in L2 English and L1 German over the course of one semester. A comprehensive set of complexity and accuracy measures (31 accuracy measures, 568 English and 717 German complexity measures) was extracted using current computational linguistic methods to operationalize the latent CAF constructs. Results show that both the L1 and L2 groups showed significant signs of development in terms of lexis and morpho-syntax, but only the L1 group exhibited development at the discourse levels as manifested by the use of more accurate cohesion and rhetoric devices, as well as more complex sentential structures. Relating complexity to accuracy development, no evidence was found for a competing relation or trade-off between complexity and accuracy. The study addresses the research gap laid out in Section 1.4.2 by making use of the comprehensive complexity analysis tool CTAP introduced in Chapter 2. It also demonstrates how to apply the complexity construct to answer SLA questions.

Related publication

This chapter is based on the following submitted manuscript:

- Chen, X., Weiß, Z., and Meurers, D. (Submitted-b). Is there a developmental trade-off between complexity and accuracy in L1 and L2 acquisition?.

4.1 Introduction

Since their emergence in the 1980s as variables for measuring language performance in Task-Based Language Teaching (TBLT), the notions of *complexity*, *accuracy*, and *fluency* (CAF) have developed into major research variables in applied linguistics (Housen et al., 2009). Besides being used as dependent variables of various task effects on written and oral language performance (e.g., Ellis and Yuan, 2004; Michel et al., 2007; Skehan, 2009; Robinson, 2011), the CAF triad has systematically been used to evaluate language proficiency (Ortega, 2003) and measure longitudinal language development (Byrnes, 2009; Polat and Kim, 2014; Vyatkina et al., 2015; De Clercq and Housen, 2017). Complexity is generally defined as the variedness, elaborateness, and inter-relatedness of language productions, accuracy as non-native-like production error rate, and fluency as native-like production speed and smoothness (Wolfe-Quintero et al., 1998; Pallotti, 2009; Housen et al., 2009, 2012).

An important concern of CAF research is how the three aspects develop, individually and in relation to each other, throughout the acquisition process. Most studies focusing on the interdependence of CAF were conducted within the TBLT framework adopting a non-developmental or static view of the constructs. They predominantly investigated the effects of various learning task factors (e.g., planned *vs.* unplanned, monologic *vs.* dialogic) on CAF performance with cross-sectional experiments (Foster and Skehan, 1996; Crookes, 1989; Tabari, 2016; Yuan and Ellis, 2003; Michel et al., 2007; Ellis and Yuan, 2004; Ahmadian and Tavakoli, 2011). Although in this case CAF serve the purpose well as descriptors of language performance attributable to different task settings, the static view of the constructs makes it very difficult to draw conclusions about how CAF develop through the learning process.

A few longitudinal studies have investigated the developmental patterns of individual CAF aspects and their relationship (e.g., Spoelman and Verspoor, 2010; Polat and Kim, 2014; Vyatkina et al., 2015; Vercellotti, 2017), but conclusions from these studies should be drawn with caution. Firstly, most studies include only a few measures to represent each CAF aspect, resulting in a narrowed view of the highly complicated and multifaceted concepts of CAF. For example, Bulté and Housen (2012) reviewed forty empirical studies on L2 complexity published between 1995 and 2008 only to find that most studies used no more than three complexity measures, although the problem has improved in the more recent studies. Secondly, non-consistent even contradicting results have been found for the developmental relationship between CAF aspects. Some studies (e.g., Wendel, 1997; Ortega, 1999; Ske-

han, 1996; Tabari, 2016; Foster and Skehan, 1996) found that CAF aspects compete against each other for the limited attentional resources, supporting Skehan's Trade-Off Hypothesis (TOH, 1998). Others, however, found that the CAF dimensions develop simultaneously as the learners' overall proficiency grows (e.g., Spoelman and Verspoor, 2010; Vercellotti, 2017), supporting the Cognition Hypothesis (CH, Robinson, 2005), which assumes that attentional resources may be directed to multiple dimensions of CAF in certain circumstances. Lastly, it is still unclear whether these results are generalizable for both first and second language acquisition since most studies up to date have focused on the L2 scenario. A partial exception is a study by Polat and Kim (2014), who compared the development of complexity and accuracy of an L2 English speaker with that of native speakers in an untutored setting. However, besides the lack of strong support to their findings because of the single-case study nature of the research, the authors did not adopt a developmental view of CAF on the L1 data.

The current study tried to address these issues by approaching CAF from a developmental point of view. A longitudinal corpus of L1 German and L2 English writings was used to investigate how complexity and accuracy as measured by an extensive set of more than 550 indexes develop during the course of a typical 15-week semester (30 contact hours and 30 hours of homework). By focusing on complexity and accuracy, we mainly target the trade-offs between the scope (complexity) and conformity (accuracy) of inter-language knowledge (Housen et al., 2009; Wolfe-Quintero et al., 1998). The longitudinal nature of the data makes it possible to account for the developmental relationship between complexity and accuracy, providing stronger evidence to support or reject the developmental hypotheses. The dataset also enables us to compare the development of CAF in L1 German and L2 English writing in a highly comparable setting—from students taking the same type of writing courses at the same university and completing the same writing tasks. To the best of our knowledge this has never been done by previous studies. Another contribution of the study is to demonstrate a new automatic tool for complexity feature extraction, the CTAP (Chen and Meurers, 2016b, see also Chapter 2) , which can save language researchers from the tedious and resource-demanding data analysis process and provides a much broader coverage of linguistic complexity measures at all levels of linguistic modeling, language use, and psycholinguistic processing complexity.

In the following sections, we will first review previous research adopting a developmental view of complexity and accuracy. This is followed by a discussion of the theoretical accounts of the relationship between complexity and accuracy

development—the TOH and the CH. Empirical evidence supporting or rejecting these hypotheses will also be reviewed before we present our research questions and the data. The discussion of our research findings will be centered around the questions of how the CAF constructs develop individually, how they interact with each other, and whether L1 and L2 development differs.

4.2 Complexity and its development

Linguistic complexity is commonly defined as the variedness and elaborateness of language production (Ellis, 2003). The definition resembles the philosophical definition of complexity as a function of the number of a system's constituent elements, the variety of these constituents, and the elaborateness of their interrelations (Rescher, 1998). Accordingly, the measurement of linguistic complexity involves quantitatively measuring the number and nature of linguistic sub-components and the interconnections between them (Bulté and Housen, 2014). The most commonly analyzed linguistic sub-components are syntax, lexicon, and—in synthetic languages—morphology. Syntactic complexity targets primarily phrasal, clausal, or sentential elements and is measured with indices such as dependent clauses per clause, complex phrases per phrase, or mean sentence length and so on (Kyle, 2016; Wolfe-Quintero et al., 1998). Lexical complexity is measured in terms of lexical *diversity*, *density*, *variation*, and *sophistication* (Bulté and Housen, 2012; Wolfe-Quintero et al., 1998), and morphological complexity is assessed with derivational, compositional, and language-specific inflectional measures (Pallotti, 2015; Bulté and Housen, 2014; Hancke et al., 2012; Reynolds, 2016; François and Fairon, 2012). Other research strands include psycholinguistic measures of discourse structure and textual cohesion such as density of connectives and co-reference constructions (Crossley et al., 2015; Graesser et al., 2004; Louwerse et al., 2004), or cognitive measures such as surprisal or cognitive processing load (Shain et al., 2016; Gibson, 2000; Vor der Brück and Hartrumpf, 2007).

Much as the construct is often used to assess the quality of language production and gauge written or spoken proficiency underlying the learner's performance (Housen et al., 2009; Ortega, 2012; Lu and Ai, 2015), the developmental view of complexity is considered the core of the phenomenon of language complexity (Ortega, 2015). Language development is associated with the increasing ability to control an ever-expanding linguistic repertoire of the target language (Ortega, 2003, 2015; Foster and Skehan, 1996). Since most CAF studies elicit data from educational contexts, which are designed to prompt learners to employ increasingly complex

language (Ortega, 2015), studies on complexity development often find that substantial exposure as well as intensive targeted instruction result in increases in complexity measure scores (e.g., Ortega, 2003; Ortega and Sinicrope, 2008; Byrnes, 2009; Lu, 2011; Vyatkina, 2012, 2013; Bulté and Housen, 2014; De Clercq and Housen, 2017; Crossley and McNamara, 2014; Mazgutova and Kormos, 2015). However, on a broader scale it should be clear that the ultimate purpose of language acquisition is not to produce increasingly complex language as an end in itself and it has often been emphasized that more complex language should not automatically be associated with more proficient or developed language ability (Bulté and Housen, 2014; Pallotti, 2009, 2015).

Mixed results have been found in previous research on the longitudinal development of complexity with regard to the areas of development and the developmental patterns. For example, in terms of the former Vyatkina (2012) found that beginning and intermediate German learners would gradually produce language that is lexically more varied and syntactically more complex with more frequent subordination as their proficiency progresses. Bulté and Housen (2014), on the other hand, found significant increase only in syntactic but not in lexical complexity from a corpus of articles by students of ESL over a period of a typical four-month semester. On the contrary, Leonard and Shea (2017) reported a significant increase only in lexical complexity as measured by the Guiraud advanced index (Guiraud, 1954) and syntactic complexity for a developmental period of three months. As for the developmental patterns, Bulté et al. (2008) found a linear progress of all lexical diversity measures in a period of three years from their French learner participants, but Vyatkina et al. (2015) reported a non-linear waxing and waning of lexical diversity development among beginning German learners. It is worth noting that these studies used different sets of complexity measures to represent the complexity subconstructs, making it difficult to compare their results.

In fact, a large number of lexical, morphological, and syntactic measures have been used in previous research: Wolfe-Quintero et al. (1998) reviewed 50 complexity measures used in studies of written language development and categorized them into the grammatical/lexical subconstructs of frequency/ratio measures. Bulté and Housen (2012) also composed an inventory of 40 linguistic complexity measures used in task-based studies. Housen (2015) even identified more than 200 features for measuring L2 complexity. However, as Bulté and Housen (2012) rightly pointed out, even though there was no shortage of complexity measures, most studies used no more than three measures to represent the complexity construct. For example, earlier research usually investigated complexity development from either the lexical (Bulté

et al., 2008; Lu, 2012; Laufer and Nation, 1995) or syntactic (Neary-Sundquist, 2017; Crossley and McNamara, 2014; Ortega, 2003; Lu, 2011; De Clercq and Housen, 2017; Lu and Ai, 2015; Vyatkina et al., 2015) perspective with a few measures. Although more recent studies (e.g., Mazgutova and Kormos, 2015; Yoon, 2017; Vercellotti, 2017; Polat and Kim, 2014; Vyatkina, 2012; Leonard and Shea, 2017) recognized the importance to approach complexity with a combination of lexical, syntactic, and/or morphological measures, the majority of them still use a few measures to represent each subconstruct. Because there is a lack of agreement on which measures 'best' represent a complexity subconstruct, different measures were used by different studies. It is thus difficult to draw conclusions on the development of complexity from them, especially when conflicting results are obtained. In the current study, the complexity construct therefore was represented by a comprehensive set of complexity measures including lexical, syntactic and morphological measures automatically extracted using computational linguistic methods. The comprehensive set of complexity measures accounts for a broader range of observable aspects of the complexity construct than any single measure or a combination of a few measures could.

4.3 Accuracy and its development

Accuracy is commonly defined as 'the degree of conformity to certain norms' (Pallotti, 2009, p. 592) and operationalized as non-native-like production error counts. It is argued to be the most straightforward construct in the CAF triad because its denotation is widely accepted. It is worth noting, however, that this position implicitly assumes a prescriptive perspective on language, which considers deviations from a language's prescribed/standard norm as errors (e.g., Housen et al., 2009, p. 463), whereas the descriptive stances on language introduce a definitional fuzziness to the notions of norms and errors, which impairs the supposed denotational clarity of accuracy. Furthermore, the explanatory power and validity of normative accuracy in SLA is sometimes questioned (Pallotti, 2009; Norris and Ortega, 2003; Wolfe-Quintero et al., 1998) based on the influential criticism by Bley-Vroman (1983), who points out the *comparative fallacy* of measuring inter-language systems against target language norms. Hence, while accuracy is currently the most agreed-upon component of the CAF triad, it is certainly not undisputed.

Notwithstanding these conceptual concerns, accuracy is commonly assessed in empirical studies on language proficiency and development. It can be operationalized as error counts, error-free units, and their normalized variants (Foster and Skehan,

1996; Wolfe-Quintero et al., 1998). Some researchers also take into consideration the nature of the errors because they found that learners of different proficiency levels produce errors of different types. For example, Taylor (1975) found that L2 learners make more over-generalization errors as their proficiency increases, while L1 interference errors are more common among less proficient learners. Others attempt to rate errors based on their severity (Kuiken and Vedder, 2008; Evans et al., 2014), which, however, has also faced some conceptual criticism with regards to the definition of weighting criteria (Pallotti, 2009). Consequently, a more comprehensive view of accuracy should include measurement of the number of (in)accurate expressions, their distribution across error types, and perhaps the severity of the errors (Polio and Shea, 2014).

Developmentally, the learner's control over form is expected to increase in the long term, resulting in the ability to produce more accurate language as appropriate to the context. However, accuracy development in shorter terms is less observable. Yoon and Polio (2017) examined essays written by ESL learners over a period of 4 months but found 'a notable lack of development in the area of accuracy' (p. 275). On the contrary, Larsen-Freeman (2006) reported an accuracy growth in the aggregated data of five Chinese learners of English over a period of six months, although their individual development trajectories varied. In a cross-sectional study Verspoor et al. (2012) found that all their accuracy measures were capable of distinguishing the students' proficiency levels. Similar results were obtained by Ishikawa (1995) who found that total words in error-free clauses and number of error-free clauses were the best measures to discriminate samples of low-proficiency writing.

The conflicting results from previous research call for further investigation into accuracy development, especially by focusing on the area of development and the developmental patterns as with the complexity construct. A developmental view on accuracy will add to our understanding of the relationship between accuracy and proficiency and shed light on the interaction between accuracy and the other CAF constructs.

In the current study, we assessed accuracy based on the diverse error annotation provided by Göpferich and Neumann (2016), who adopted a comprehensive error classification scheme by classifying errors into five categories: formal, lexical, grammatical, text-linguistic, and other errors. The classification scheme allows us to capture a more elaborate picture of accuracy development than most previous developmental studies. The accuracy measures listed in Appendix B were extracted from the data.

4.4 The interaction between the CAF constructs

The inter-relatedness of the CAF components is commonly discussed in terms of trade-off effects which describe the prioritization of one dimension of language performance at the expense of the others. Two theoretical accounts dominate the discussion: Skehan's TOH (1998) and Robinson's CH (2005).

The TOH, also known as the Limited Attentional Capacity Model, describes the inter-relatedness of CAF dimensions in terms of a competition for limited attentional resources: It assumes that all dimensions of language performance draw from the same pool of limited attentional resources and that prioritization of one of the CAF components leaves the competing dimensions with diminished resources, thus resulting in poorer performance in these dimensions (Skehan, 2009). The primary trade-off effects may be observed between focus on form (complexity, accuracy) and focus on meaning (fluency) (VanPatten, 1990). If form is prioritized, a secondary trade-off between the scope (complexity) and conformity to form (accuracy) is assumed to take place. In a developmental setting, trade-off effects are assumed to result in the prioritization of a single area of language performance, hindering progress in the other dimensions (Kuiken and Vedder, 2007) until the prioritized area becomes automatized enough to release the previously allocated attentional resources for process and storage of the hindered areas (Göpferich and Neumann, 2016; McCutchen, 1996).

The CH, also known as the *Triadic Componential Framework*, takes a contrasting stand on cognitive limitations on language performance. It rejects the assumption of a single limited attentional resource pool for all CAF components but favors a multiple-resource interferential account. In this purely task-based framework, the central components mediating the inter-relatedness of the CAF dimensions are *task complexity* (cognitive factors), *task conditions* (interactional factors), and *task difficulty* (learner factors) (Robinson, 2001). The CH distinguishes *resource-directing* (cognitive/conceptual) and *resource-depleting* (performative/procedural) dimensions of task complexity. It stipulates that increases in task complexity along the former dimension enhance attention to input and output and thus promote the development of accuracy and complexity by facilitating noticing of relevant structures (Robinson and Gilabert, 2007). Increased task complexity along resource-depleting dimensions will impede access to the current repertoire of L2 knowledge due to loss of control during central processing and interferences during resource allocation, resulting in decreased language performance (Robinson, 2003; Vercellotti, 2017). This may yield trade-offs between CA and F caused by involuntary attention shifts but not between C and A (Robinson, 2003, p. 645). Crucially, however, it

is assumed that simultaneous prioritization of complexity and accuracy or all three CAF dimensions is possible, whereas the TOH rejects such a parallel enhancement. In essence, the TOH and the CH both predict decreased complexity and accuracy for increased performative demands, but the later further postulates that increased cognitive task demands would result in increased language performance in all CAF dimensions. These opposing predictions have been exploited by developmental studies on trade-off effects, which do not modify tasks. Although the CH is tied to task effects, the simultaneous developmental progress of accuracy and complexity is typically taken to speak against Skehan's secondary trade-off and in favor of the CH.

Although both hypotheses make partially contradicting predictions about trade-off, findings across empirical studies showed mixed results (Yoon and Polio, 2017). The TOH finds support from research in first language acquisition where emergent writers spend the majority of their attentional resources in translating ideas into text, producing less complex and less accurate language (McCutchen, 1996), resulting in trade-off between CA and F as Skehan predicts. Various studies in second language acquisition also found trade-offs between different CAF constructs under different task setups. For example, Yuan and Ellis (2003) found that pre-task planning is effective in increasing the complexity and accuracy of learner writing at the expense of fluency (i.e., trade-off between CA and F). Skehan and Foster (1997) and Ferraris (2012) both found competing relationship between the complexity and accuracy constructs. However, counter evidence has also been reported in previous research. After analyzing a longitudinal dataset of L2 speech monologues by 66 participants of different L1 backgrounds in terms of their CAF development over time, Vercellotti (2017) concluded that their results 'do not support the supposition of trade-off effects' (p. 91). Robinson (1995) also found that cognitively more difficult tasks are likely to result in increase in both complexity and accuracy, disproving Skehan's account of the developmental relationship between the CAF constructs.

Empirical support to the CH has been found in a number of studies too, though. For example, Kuiken and Vedder (2007) found that manipulation of task complexity led to significant increase in both accuracy and lexical variety of L2 writing. Michel et al. (2007) showed that more complex tasks generated more accurate though less fluent speech in both monologic and dialogic conditions, while the dialogic tasks triggered increase in both accuracy and fluency but decrease in complexity. Robinson (2011) reviewed a number of studies involving various task complexity variables (e.g., ± here and now, ± few elements, ± causal reasoning) and found support to the CH. One problem with this line of research is that it mainly focuses on the effects of tasks on CAF measures but sheds little light on the developmental interaction between

the CAF constructs.

In light of the above reviews, the current study seeks answers to the following research questions:

1. How do complexity and accuracy develop longitudinally in the writing of subjects writing in the second and in their native language over the period of one semester?

2. What is the relationship between complexity and accuracy development, specifically, is there evidence for competition?

4.5 Method

A longitudinal corpus consisting of German (L1) and English (L2) essays written by students over a period of one semester was used to answer the research questions. Operationalized complexity measures were extracted from the corpus automatically using current computational linguistic methods. The accuracy measures were calculated from the manual error annotation constructed in Göpferich and Neumann (2016). We used SEM methods for data analysis because it is especially useful for research questions that (i) involve complex latent constructs measurable by a number of observable indicators (often with errors), (ii) need to tackle a 'system' of relationships, and (iii) need to investigate the direct and indirect effects of predictor variables (Sturgis, 2016). SEM is also usable with longitudinal data (Barkaoui, 2014).

4.5.1 The corpus

The corpus used in the current study was originally collected by Göpferich and Neumann (2016) in a project to evaluate the development of students' L1 and L2 writing competences after one semester of instruction. It consists of argumentative essays written by students enrolled in a general writing course offered in a German university. The course was offered in both German and English which were the students' L1 and L2 respectively. Students could choose to take the course in either language and were required to complete the essays in the course language. Two essays were written by each student: one at the beginning of the semester and the other at the end. The writing assignments were timed (90 minutes each) argumentative essays with a length limit of 250 to 350 words for each essay on topics assigned by the course instructor. The students were free to choose one from three different topics in each assignment and the topics were different between the two

assignments. The invariant task setting of essay solicitation makes the data especially suited for answering our research questions since the data is not subject to varying task-effects like many previous longitudinal studies (Yoon and Polio, 2017; Tracy-Ventura and Myles, 2015). All the essays were manually error-annotated in a discursive consensual assessment procedure using a comprehensive error classification scheme and holistically evaluated in terms of the argumentative rigor of the essay by three raters. We calculated the accuracy measures based on the error annotations from the original corpus. Table 4.1 summarizes the profile of the corpus.

	German			English		
	Begin	End	T*	Begin	End	T*
# essays	50	28		66	41	
Mean length (words)	337.96	346.11	$-.74^{\dagger}$	372.77	367.66	$.41^{\dagger}$
SD length	54.70	40.95		66.20	59.70	

*: two-tailed independent samples T-test; †: non-significant

Table 4.1: Corpus profile

The mismatch between the number of essays collected from semester begin and end was caused by student dropout from the course. In order to make full use of the data, the missing data were estimated with an maximum likelihood algorithm from the existing data when the SEM model was fitted.

4.5.2 Complexity and accuracy measures

A large number of measures have been used to quantify complexity and accuracy for various research purposes. For example, complexity has been operationalized on multiple linguistic levels including lexical, morphological, and syntactic levels to account for proficiency development (Bulté and Housen, 2014; Ellis and Barkhuizen, 2005; Mazgutova and Kormos, 2015), text readability (Benjamin, 2012; Collins-Thompson, 2014; Vajjala and Meurers, 2012), task effects (Robinson, 2011; Foster and Skehan, 1996; Tabari, 2016) and so on. Over 200 measures have been used in previous studies. We have developed analytical tools capable of extracting more than 550 complexity measures from English texts (Appendix A). Accuracy, operationalized mainly as error counts and sometimes as holistic scales (Polio and Shea, 2014) has also been used to investigate task effects on language production (Ellis and Yuan, 2004; Kuiken and Vedder, 2008; Kormos, 2011), the effects of feedback types (Evans et al., 2010; Bitchener and Knoch, 2010; Chandler, 2003), proficiency development (Storch, 2009) and so on. Polio and Shea (2014) reviewed 44 accuracy measures used in recent accuracy studies.

Different operationalizations of the complexity and accuracy constructs by previous studies make it difficult to compare results and draw consistent conclusions across studies. Therefore in the current study, we included a comprehensive set of complexity and accuracy measures that have been used in earlier studies: 31 accuracy measures (Appendix B), 568 English complexity measures (Appendix A) and 717 German complexity measures (cf. Weiß, 2017). The extensive set of measures allows us to construct a more complete picture of the development of complexity and accuracy in writing. Calculating such a large number of measures from the texts is a resource demanding task. Fortunately, automatic tools such as CohMetrix (McNamara et al., 2014), CTAP (Chen and Meurers, 2016b, see also Chapter 2), and the L2 Lexical/Syntactic Complexity Analyzers (Lu, 2010, 2012) that make use of natural language processing technologies have been developed to aid the analysis. We used the web-based CTAP system for extracting the complexity measures. The accuracy measures were calculated from the manual error annotations conducted by Göpferich and Neumann (2016) who classified learner errors into six main categories and 28 subcategories (ibid, Table 1, pp. 115–118).

4.5.3 Statistical procedures

To answer the first research question about how accuracy and complexity developed over the period of a semester, paired sample T-tests were run on each measure to compare their means. Data instance with missing values were removed pair-wisely. Given the number of measures that we tested for significance, our analysis runs the risk of sporadically showing significant results by chance (alpha error inflation). In order to address this issue, we grouped our measures into nine theoretically defined feature sets, which we set a priori. We only consider feature groups that show a development in at least 10% of their measures, because we consider this accumulation of significant results as evidence for a significant effect beyond chance. Table 4.2 lists the grouping of the complexity measures.

To answer the second research question, we adopted an SEM procedure to analyze the data. SEM is a statistical method for testing a hypothesized network of relationships among some latent and observed variables based on the variances/covariances between these variables (see, for example, Schumacker and Lomax, 2010, for an introduction). An SEM model usually consists of two components: a measurement model and a structural model. The former hypothesizes and tests the relationship between the latent constructs and the observed variables manifesting them. The latter investigates the relationships between the latent constructs. Our research questions were mainly about the development of the complexity and

Category	# measures English	# measures German	Explanation
Lexical density	96	49	Raw counts of lexical components categorized by POS in tokens or types (unique tokens) as well as the ratio of these components to all tokens/types. The POS abbreviations in the measure names follow the Penn Treebank POS tag set (Taylor, 2003).
Lexical variation	43	68	The variedness of lexical use as measured by TTR and their normalized variants to account for text length effect.
Lexical sophistication	360	125	Sophistication of lexical choice as measured by the frequency of words in representative language samples, i.e., large corpus of normed language use. Multiple lists (BNC, SUBTLEXus, NGSL) and frequency indices (normalized frequency counts, contextual diversity and so on) are used in the measures. Measures of semantic relations are also categorized as lexical sophistication indices.
Syntactic density	30	93	Raw counts of syntactic components in the phrasal, clausal and sentential levels. Again, Penn Treebank's syntactic tagset (Taylor, 2003) was used to categorize the components.
Syntactic complexity/ratio	17	155	Normalized ratio of syntactic components to number of clauses, sentences, and T-units.
Morphology	NA	85	Raw counts and ratios of morphological language properties including inflection, derivation, and composition based on POS tags, morphological analyses, and compound analyses following suggestions by Hancke et al. (2012).
Cognitive processing	NA	47	Raw counts and ratios of cognitive integration costs assessing cognitive processing load based on Gibson (2000)'s Dependency Locality Theory and the operationalizations by Shain et al. (2016).
Cohesion	22	95	Use of cohesive devices in subsequent sentences (local) or across all sentences (global) in a text as well as syntactic similarity between sentences as measured by parse tree edit distance.

NA: no measures in that category

Table 4.2: Summary of complexity measures and their categories

accuracy constructs and their relationship. Therefore the focus of our analysis is on the structural model.

Figure 4.1 shows the model we used to fit the data. The latent constructs of complexity and accuracy were measured twice at the beginning and end of a writing course, represented by the four ellipses in the figure. Each latent construct was manifested by three parcels of measures, represented as squares named `cpxx` and `apxx`. Variable parceling (also known as item parceling) was used because of the abundance of the complexity and accuracy measures and small sample size of the dataset. Item parceling is argued to bring a number of benefits to SEM such as alleviating psychometric problem, improving model efficiency, and remedying small sample size (Matsunaga, 2008). Following Matsunaga's suggestions, we randomly assigned the observed complexity variables into three parcels and the accuracy variables into another three parcels. The variables that went into each parcel were the same across the two measurement points—semester begin and end. The *structural model* in Figure 4.1 is represented by the double-headed bold arrows linking the latent constructs, while the *measurement model* is represented by the single-headed light arrows pointing from the latent constructs to the parceled indicators. The circles represent the error terms or disturbances of the measurement.

4.5.4 Computational tools

The data analysis was done in R (R Core Team, 2015) and the SEM model was fitted with the `lavaan` package (version 0.5-23.1097, Rosseel, 2012), a free open-source package for **la**tent **va**riable **an**alysis. We used the Confirmatory Factor Analysis (CFA) function with the Maximum Likelihood estimator to fit the model. Missing values were estimated with case-wise (or 'full information') maximum likelihood estimation provided by `lavaan`.

4.6 Results

4.6.1 Results for RQ 1

Since a data-driven approach was adopted for the research, a comprehensive set of complexity and accuracy measures (Appendices A and B) were tested to investigate the areas of development over the semester. Table 4.3 summarizes the number of different types of measures where significant changes have been observed between the beginning and the end of the semester for both the English and German groups. It is worth noting that the length of the participants' essays were not significantly

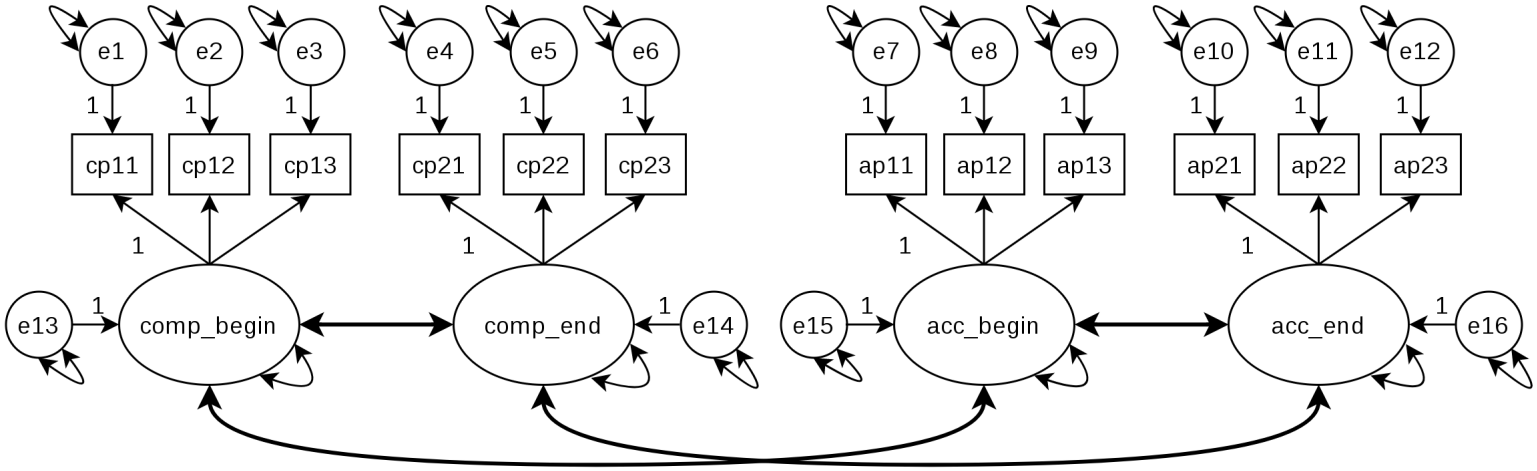

Figure 4.1: The structure equation model used in the study.

Note: Light arrows: the measurement model; bold arrows: the structural model; ellipses: latent constructs; squares: parceled indicators; single-headed arrows: directed relationship: double-head arrows: (co)variances or correlation; e: error; 1: fixed regression coefficients; begin/end: data from semester begin/end

different between the begin and the end of the semester (see Table 4.1). This makes the density measures that are not normalized for text length comparable across measurement points.

Type of Measures	Sig./Total Measures (Ratio) English	German
Accuracy	3/31 (9.6%)	5/31 (16.1%)
Complexity		
Lexical density	20/96 (20.8%)	11/49 (22.4%)
Lexical variation	12/43 (27.9%)	3/68 (4.41%)
Lexical sophistication	92/360 (25.6%)	32/125 (25.6%)
Syntactic density	8/30 (26.7%)	13/93 (14.0%)
Syntactic complexity/ratios	3/17 (17.6%)	1/155 (0.65%)
Morphology	NA	14/85 (16.5%)
Cognitive processing	NA	3/47 (6.38%)
Cohesion	0/22 (0.0%)	13/95 (13.7%)
Total	138/599 (23.0%)	95/748 (12.7%)

Table 4.3: Ratio of measures showing significant changes between semester begin and end

The results showed increased accuracy of student essays in terms of normalized number of errors, as well as number of form and punctuation errors for the English group. As for the students taking writing course in their L1 German, the accuracy of their writing also improved, but in different areas from their English counterparts. While L2 development occurred on the formal levels of language use, L1 accuracy development happened mainly on the discourse level which was manifested by less implicitness, repetition, rhetoric and coherence errors.

The results also suggest development in the area of lexical sophistication as manifested by the decreased mean frequency values towards the end of the semester for both the L1 and L2 groups. Decreased frequency of words chosen by the students means increased usage of less-frequent words which are often acquired later or used by more advanced learners, hence higher lexical sophistication.

However, for most lexical diversity measures (TTR and its variants) that showed significant changes, lower diversity values were observable by the end of the semester for the L2 group. This is also reflected in the lexical density measures where the counts of certain lexical types (e.g., number of word types and number of personal pronoun types) decreased significantly. The token count measures in the lexical density category are not interpretable because increase/decrease in certain token types would mean decrease/increase in others, given that the length of the writings did not differ significantly. Development of lexical variation was barely observable

from the L1 German group.

Syntactically, development can be observed on the sentential and phrasal levels for both groups. By the end of the semester, students were able to write longer and structurally more diverse clauses in English. Yet, the total number of clauses, complex T-units, dependent clauses, adjectives, nouns, and verbs decreased. In contrast, the students taking the German writing course used more clauses and dependent clauses at the end of the semester as well as more complex noun phrases and postnominal modification.

We also included morphological complexity measures from the German data, since this domain has shown to be highly relevant for languages with rich-morphology like German (cf. Hancke et al., 2012; François and Fairon, 2012; Vor der Brück et al., 2008). We found development in 16.5% of the measures. These predominantly include certain types of nominalizations as well as the amount and depth of compounds.

None of the cohesion measures showed significant changes over the semester for the English group. Conversely, 13 out of the 95 (13.7%) German cohesion measures showed significant development over the semester. This finding echoes the results of accuracy development that L1 development mainly happened on the discourse levels, while L2 development was more on the lexical levels.

To sum up, these results showed that after a semester's instruction, the participants were able to use more complex words and syntactic structures in their writings while producing fewer mistakes. However, in terms of the areas of development, the L1 and L2 groups showed different patterns. While the L2 group was still mastering the usage of the linguistic forms of the target language, the L1 group was also developing abilities on the discourse levels to write articles that are more cohesive, explicit and with more accurate rhetoric effects.

4.6.2 Results for RQ 2

The SEM model in Figure 4.1 was fitted to both the English and German data. However, the German model failed to converge because of the small sample size—the generally agreed-on ratio of sample size to number of free parameters to be estimated is 10:1 (Schreiber et al., 2006; Bentler and Yuan, 1999), although some researchers suggested a ratio of 20:1 or higher, depending on factors such as model quality, data distribution, and missing values (Kline, 2011). As a result, we report only the results from the English group in this study. Table 4.4 lists the correlation coefficients of the indicator parcels for CFA and SEM. It can be seen from the table that the complexity indicator parcels correlated highly with each other at both

the semester begin and end measurement points. The accuracy indicator parcel correlations were weaker but still significant. These high correlations among the complexity indicator parcels suggest that the parceled indicators were measuring the same latent construct of complexity. The same applies to accuracy parcels. On the other hand, the low correlations between the complexity and accuracy parcels indicate that the latent constructs of complexity and accuracy are distinct.

The SEM model fit measures are summarized in Table 4.5. The results suggest a good fit of our hypothesized model according to the fit measure standards recommended by Byrne (2001) and Dion (2008). No post-hoc modifications to the model were conducted because of the good fit of the current model. Estimated parameters of the model are listed in Table 4.6. The standardized parameters are also shown in the graphic representation of the model in Figure 4.2.

The SEM results confirmed the validity of our measurement model—significantly high factor loadings were found for all parceled indicators. The standardized factor loadings range from .603 to .988. In terms of the structural model as represented by the bold arrows connecting the latent ellipses in Figure 4.2, we were able to find significant correlations between the same constructs across time (begin and end of semester). The standardized correlation coefficients were .312 and .470 for complexity and accuracy respectively. These results suggest that the complexity and accuracy constructs were consistent hence the indicators were robust manifestation of the interested constructs. However, insignificantly low correlations were found between complexity and accuracy at both measurement points (.228 for semester begin and .062 for semester end). In other words, the level of complexity of student writings does not predict their accuracy, and *vice versa*. Consequently, the results of the structural model do not support the claim that trade-off occurs between complexity and accuracy.

4.7 Discussion

Our first research question was about how accuracy and complexity of student writings developed over the course of a semester. The results summary in Table 4.3 showed different developmental patterns between the L1 and L2 in both accuracy and complexity, likely due to a combination of different instructional course topics (Göpferich and Neumann, 2016, pp. 113–114) and proficiency differences between the students' L1 and L2, hence different developmental priorities. The second research question was about the developmental interrelation between accuracy and complexity. They were found to develop simultaneously and not mutually suppres-

	1	2	3	4	5	6	7	8	9	10	11	12
1. Complexity Begin Parcel 1	1.00											
2. Complexity Begin Parcel 2	.94	1.00										
3. Complexity Begin Parcel 3	.90	.94	1.00									
4. Complexity End Parcel 1	.39	.31	.28	1.00								
5. Complexity End Parcel 2	.39	.31	.29	.92	1.00							
6. Complexity End Parcel 3	.31	.22	.22	.93	.93	1.00						
7. Accuracy Begin Parcel 1	.20	.20	.17	-.12	-.07	-.17	1.00					
8. Accuracy Begin Parcel 2	.10	.12	.08	-.02	-.06	-.10	.56	1.00				
9. Accuracy Begin Parcel 3	.18	.23	.21	-.01	.03	-.03	.64	.50	1.00			
10. Accuracy End Parcel 1	.07	.04	-.00	.11	.12	.05	.33	.21	.21	1.00		
11. Accuracy End Parcel 2	.02	-.01	-.01	-.03	.04	-.08	.21	.17	.04	.54	1.00	
12. Accuracy End Parcel 3	-.02	-.07	-.05	.05	.08	-.01	.44	.25	.23	.72	.52	1.00

Table 4.4: Correlations of parceled indicators for CFA and SEM from the English subset

	Value	Good fit recommendation
Ratio of χ^2/df	1.051	≤ 3, close to 1
P-value (χ^2)	.378	$> .05$
RMSEA	.028	$\leq .05$
SRMR	.044	$\leq .08$
CFI	.997	$> .95$
TLI	.995	approaches 1

Table 4.5: Fit measures of the SEM model

	Latent construct	β	B	SE
Complexity Begin Parcel 1 (cp11)	Complexity Begin	.956	1.000	
Complexity Begin Parcel 2 (cp12)	Complexity Begin	.988	1.045	.047
Complexity Begin Parcel 3 (cp13)	Complexity Begin	.947	1.021	.058
Complexity End Parcel 1 (cp21)	Complexity End	.962	1.000	
Complexity End Parcel 2 (cp22)	Complexity End	.963	1.052	.054
Complexity End Parcel 3 (cp23)	Complexity End	.962	.099	.051
Accuracy Begin Parcel 1 (ap11)	Accuracy Begin	.921	1.000	
Accuracy Begin Parcel 2 (ap12)	Accuracy Begin	.616	.599	.135
Accuracy Begin Parcel 3 (ap13)	Accuracy Begin	.701	.808	.158
Accuracy End Parcel 1 (ap21)	Accuracy End	.815	1.000	
Accuracy End Parcel 2 (ap22)	Accuracy End	.603	.533	.107
Accuracy End Parcel 3 (ap23)	Accuracy End	.887	1.133	.201
Complexity Begin	Complexity End	.312	388.445	165.612
Accuracy Begin	Accuracy End	.470	4.932	1.693
Complexity Begin	Accuracy Begin	.228	37.453	22.325
Complexity End	Accuracy End	.062	4.928	10.790
Complexity Begin	Accuracy End	-.028	-2.846	13.958
Complexity End	Accuracy Begin	-.115	-14.544	17.241

Table 4.6: Standardized and unstandardized parameters for SEM model fitted to the English data

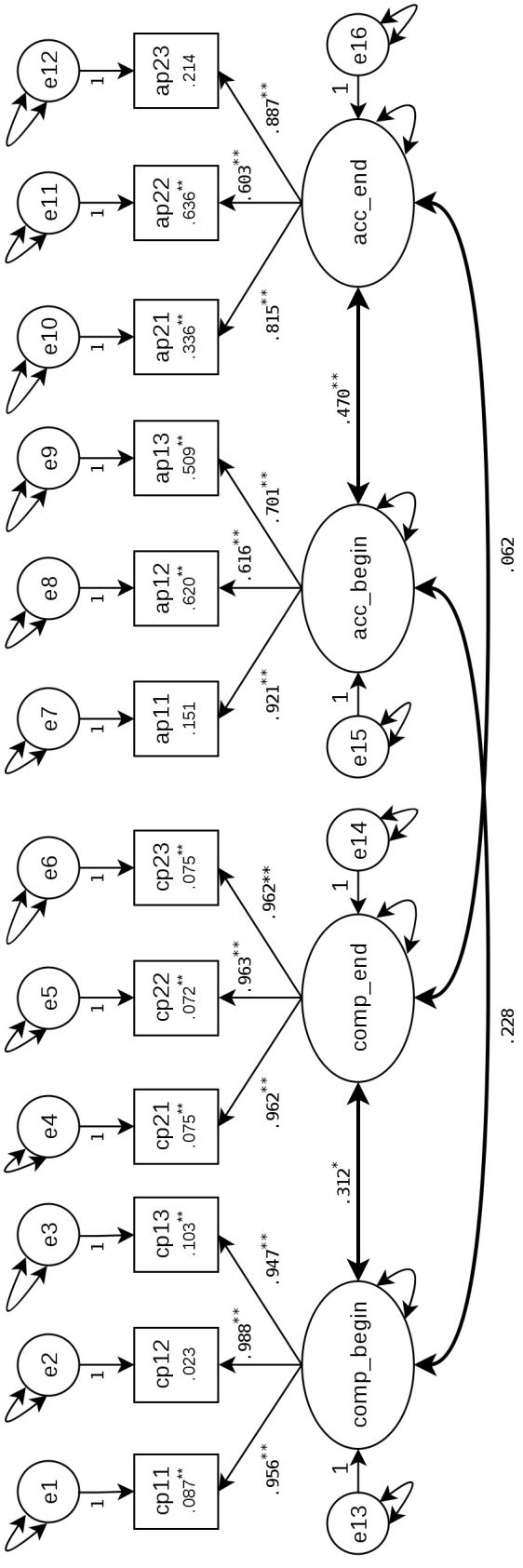

Figure 4.2: SEM model with standardized parameter estimates

** : $p < .01$; * : $p < .05$

sive, rejecting the TOH but supporting the CH.

4.7.1 Accuracy development

Accuracy development was found in the L2 English group in three measures: errors per 100 words, number of punctuation errors, and number of form errors. For the German L1 group, general number of errors per 100 words also decreased, as well as numbers of implicitness, repetition, rhetoric, and cohesion errors. These findings contradict results from Yoon and Polio (2017), who failed to find notable improvement in accuracy within a comparable period of time (four months) and with similar participants to the current study. However, our results corroborate findings of other studies. For example, Vercellotti (2017) found a linear growth in accuracy as measured by error-free clauses over time. Larsen-Freeman (2006) also observed accuracy development in terms of error-free T-units over a six-month period, although a similar measure of error-free sentences did not show any improvement in both our L1 and L2 data. Polio and Shea (2014) also investigated detailed error types but found improvement only in preposition errors over a semester.

One possible reason for the deviation of our findings from some of the previous research is that our data were collected in a different learning environment. Most previous studies—including those cited in the previous paragraph—were conducted under the ESL environment (predominantly in the US), whereas the current study was done in an English as a Foreign Language (EFL) setting. Learners in an ESL environment tend to focus more on fostering smooth communication between themselves and the native speakers they come across in their everyday life, hence prioritizing fluency but less accuracy. For example, in spite of the lack of development in accuracy from their students, Yoon and Polio (2017) found a significant time effect on fluency, which suggested that their ESL participants wrote longer essays within the same time as they developed their proficiency. Another reason for the deviation is probably due to instructional effects. Part of the instructions the participants received was on punctuation rules and refreshment of grammatical knowledge for both the English and German groups, thus improvement in the instructed areas was not out of expectation. Last but not least, the comprehensive set of accuracy measures also enabled us to create a fuller account of accuracy development. Yoon and Polio (2017) used accuracy measures on four detailed linguistic levels (syntactic, morphological, preposition, and spelling errors per 100 words) to account for accuracy development and did not find any. However, their finding does not rule out the development of accuracy in the other unmeasured areas.

In comparison to L2 accuracy development, L1 development occurred in different

areas. Whereas L2 participants mainly developed in more local form-related areas, L1 accuracy development happened mainly in the areas of meaning and more global areas, such as more accurate rhetoric effects and text cohesion. In fact, all the four types of specific errors where development was observed in the L1 data were classified into the text-level (text-linguistic) error category by Göpferich and Neumann (2016, p. 118). It is understood that native speakers make less form errors because they have a higher level of automatization in grammar and lexical usage so they can focus more on the more global textual levels. But for L2 learners, in the short term, textual-level accuracy improvement is less observable because they still need to deal with form accuracy.

In summary, the accuracy results from the current study confirmed the longitudinal development of accuracy over a short period of time as some previous studies showed. This finding is also supported by previous studies with cross-sectional data (e.g. Verspoor et al., 2012; Ishikawa, 1995) which found that accuracy measures were able to discriminate learners' proficiency levels. Furthermore, we were also able to pinpoint the areas of accuracy development in L1 and L2. Native speakers taking compensatory writing courses in the university level tend to develop accuracy in the textual level, while L2 learners developed more on the lexical and formal levels.

4.7.2 Complexity development

In terms of complexity, development was found in lexical and syntactic aspects for both the L1 and L2 groups. While development on the textual level as manifested by cohesion measures was unobservable from the L2 group, the L1 group showed significant development in a few textual measures (e.g., local noun overlap, concessive connectives per sentence, transitional probabilities of grammatical roles from object to adjuncts in adjacent sentences, etc.). These results provide support to some previous findings on complexity development (e.g., Vyatkina, 2012; Leonard and Shea, 2017; Bulté et al., 2008), but also contradict some others (e.g., Bulté and Housen, 2014; Vyatkina, 2015). Vyatkina (2012) found significant correlations between five out of the six complexity measures used in her study and time from the writings of a cohort of students taking intermediate- and beginning-level L2 German courses in four sequential 16-week-long semesters. She found that the participants were able to produce longer sentences, use more coordinating and subordinating conjunctions, and write lexically more diversified essays.

Our results partially corroborate Vyatkina's (2012) findings regarding clausal development. The participants in the L2 group were able to write longer clauses and sentences as well as sentences with more varied structures as suggested by the

increased standard deviation of syntactic edit distance of parse trees. This finding corroborates those by Bulté and Housen (2014), who saw development in their ESL students in the syntactic aspects over a four-month period.

Our L1 group produced significantly more dependent clauses as in Vyatkina (2015), who found that her beginning and intermediate L2 German learners used more subordinate clauses and syntactically more complex sentences as their proficiency improved. Vyatkina observed the development trend with data collected over a 4-semester span. The current study showed that syntactic complexity development is also observable in a shorter term for L1 productions. Interestingly, the participants in the English group used less dependent clauses and fewer complex T-units at the end the semester than in the beginning. One relevant aspect might be that L2 syntactic development would require a longer period of instruction than L1 development. For example, Vyatkina (2015) observed syntactic development in L2 over a period of four semesters. However, it is particularly interesting that subordination decreased for the L2 English group, while it increased in the L1 group. An increased hypotactic clausal structure is a typical element of German academic language (Kretzenbacher, 1991; Beneš, 1976; Panther, 1981). The increase in subordination in the L1 texts is thus an expected index of their progressing development. For the L2 texts, it seems reasonable to assume, that German L2 writers of English initially overuse subordination. Our results illustrate how students reduce this inappropriate L1 transfer in the course of instruction. Identifying such differences between German and English writing style is also an explicit objective of the L2 writing courses (Göpferich and Neumann, 2016, pp. 113–114).

Other studies also observed short- and long-term development in lexical diversity as measured by the various TTR indices (e.g., Leonard and Shea, 2017; Bulté et al., 2008), while the current study found the opposite. For example, Leonard and Shea (2017) found that the lexical diversity (measured with the Guiraud index, a type of TTR measure) of their Spanish learners' had a significant increase after studying in a Spanish speaking country for three months. On the contrary, the TTR measures including the Guiraud index in the current study all decreased uniformly towards the end of the semester. This may be due to the relatively short span of time between the measurements, as the same phenomenon was also observed by Vercellotti (2017), who found 'a slight decline and followed by steeper increase over time' (ibid. p. 103) in lexical diversity, but over a period of up to 10 months. Bulté and Housen (2014) did not find significant correlation between their lexical diversity measures and time in ESL students' writing over a four month period either.

We also found that L2 and L1 complexity development was also manifested

in lexical sophistication with students using less frequent words in their writings. While vocabulary development is listed in the English writing curriculum as an instructional objective, it is not an explicit objective in the German writing course. Still students taking the German course developed in this regard, possibly because of increase exposure and practice in academic language through the course, which resulted in the transformation of passive to active knowledge of the L1.

Our results also showed a clear development of nominal writing style for the German data. This is a major characteristic of German academic language (Hennig and Niemann, 2013; Kretzenbacher, 1991; Beneš, 1976). Although nominal writing style was not a component of the course curriculum, we found significantly more complex noun phrases and postnominal modifications as well as increases of noun compounds, compound depths, and certain nominalizations. All of these contribute to a writing style that organizes information primarily within the nominal domain. Again, we believe this is because of increased exposure and practice as was the case in vocabulary development.

In terms of textual cohesion, the German group exhibited significant improvement while the English group did not show any development. In particular, for the German group, our results showed decreased use of connectives and increased use of implicit cohesion devices such as transitions of grammatical roles across sentences. These results were in line with the findings of accuracy development where the L1 German group showed exclusive development in discourse measures. Our findings also corroborate findings from previous research on cohesive development: more proficient writers tend to employ fewer explicit cohesion markers, but rely more on lexical coreference (Crossley and McNamara, 2012; Crossley et al., 2014; McNamara et al., 2009). However, the English data did not show any cohesive development despite the fact that argumentative structure, text coherence and cohesion, and in particular the proper use of logical connectives were listed as instructional foci of the writing course (Göpferich and Neumann, 2016, 113-114). The lack of development in this domain as exhibited in the English data suggests that cohesive development was not a developmental priority of the L2 group, probably due to the lack of readiness to develop in this respect because of the higher cognitive demand involved in using the cohesive devices which are build upon proficient usage of lexical and syntactic components of the language. Complexity development on the textual level, as manifested by more frequent use of cohesion and coherence devices (Ghasemi, 2013; Crossley et al., 2016), requires better mastery of the lower level abilities. L2 writing has been shown to involve higher cognitive demand in the lower levels of lexis and grammar than L1 writing, thus creating negative effects on the higher-level

performance like textual cohesion (Silva, 1992; Cumming, 2001).

Synthesizing results from both the current study and those from previous research, it is clear that complexity does develop both in the short term and in the long run. L2 complexity development occurs mainly on the lexical and sentential levels, probably because of instructional focus and the learners' lower proficiency in the L2. Increased complexity in terms of higher cohesion was only observed in the L1 data, although both the German and English courses focused on fostering more cohesive writing. The L1 group further exhibited development in the phrasal and clausal domain as shown by increased use of constructions that are characteristic for academic writing. Lexical development in L1 was restricted to the use of more sophisticated vocabulary. Overall, our findings clearly show how L1 and L2 language development depends on the interplay between instructional focus, the learner's developmental potential, and certain target language characteristics. By comparing the results of the current study with those of previous research, the advantage of using a more comprehensive set of complexity measures is obvious. The set of measures used in the current study is far from complete though it covers an extended proportion of measures used in earlier research. This comprehensive set of measures enables researchers to create a more complete picture of complexity development in both L1 and L2. Furthermore, it makes our findings more solidly grounded because it better avoids the 'absence of evidence is not evidence of absence' problem.

4.7.3 A complexity-accuracy trade-off?

As regard to the second research question on the longitudinal interrelationship between accuracy and complexity, the structural component of SEM model in Figure 4.2 provides us with a means to answer the question. The good fit of the model (see Table 4.5) confirmed the validity of the complexity and accuracy measurements and the potential network of relations among the two constructs. The lack of predictivity of complexity to accuracy, or *vice versa* as suggested by the insignificantly low correlation coefficients between them at the both measurement points (.228 and .062 at semester begin and end respectively) suggest that the two constructs developed independently. Whereas the current analysis did not provide evidence to the relationship between the development of form (complexity and accuracy) and meaning (fluency) because of the lack of fluency measures for written data, our results suggested that the hypothesis of the existence of trade-offs between complexity and accuracy could not be supported. This finding contradicts those of Skehan and Foster (1997) and Ferraris (2012) but corroborates findings of Vercellotti (2017), Robinson (1995) and Kuiken and Vedder (2007). Different from

research on TBLT where researchers usually investigate the effects of different task factors on the learners' language performance, the current study did not manipulate task factors because the student writings were collected with the same writing task. Consequently, the difference in complexity and accuracy performance is attributable to proficiency development. Development in both accuracy and complexity was observed in the data, disconfirming the TOH but in line with Vercellotti's (2017) findings, which were based on spoken instead of written data. The results from the current study partly support Robinson's (2011) hypothesis on the possibility of fostering simultaneous development in both accuracy and complexity if the learners are given tasks with increased complexity along the resource-directing dimension, but we also show that the two constructs can develop without manipulating the task factors. As a result, the findings shed more light on language development as regard to the complexity and accuracy constructs, rather than on task effects on language performance like most other studies did.

4.8 Summary

Complexity and accuracy and their trade-offs have traditionally been investigated from a static perspective as dependent variables of task factors. Recent years have seen increasing use of these constructs in research on language development in both L1 and L2 settings with both cross-sectional and longitudinal data. In the current research, we are interested in how complexity and accuracy develop over a relatively short period of time for language acquisition and whether their development differs in L1 and L2. We also tested the interrelations between the two constructs for the purpose of finding out whether there are trade-offs in their development as predicted by some theories. Results show that both the L1 and L2 groups showed signs of development in the lower linguistic levels of lexis and syntax, but only the L1 group demonstrated development on the higher textual level with increased cohesion in writing. Both groups developed in terms of accuracy and again, the improvement of the L2 group is on the lower levels of linguistic and typographic forms, while the L1 group improved mainly on the higher levels of rhetorics, cohesion and explicitness of their language. No trade-off was found between accuracy and complexity, meaning that the development of one aspect does not necessarily mean the suppression of the other.

The methods adopted by and results obtained from the current research have important implications for CAF research and practical language teaching. Firstly, the comprehensive set of complexity and accuracy measures used in the study enables

us to create a more complete picture of the longitudinal development of the constructs under investigation, which are multidimensional and multifaceted concepts. Secondly, together with similar findings from other studies (e.g., Robinson, 2011; Larsen-Freeman, 2006; Vercellotti, 2017), we would suggest that it is not necessary for language instructors to prioritize any one aspect—complexity and accuracy can develop simultaneously. Thirdly, the different developmental areas between the L1 and L2 groups due to different proficiency levels in the two languages suggest that the focus of instruction should be adjusted to the learner's abilities. Last but not least, the automatic language analysis system CTAP (Chen and Meurers, 2016b) can be used not only by language researchers to study language development, but also by language teachers to identify the areas where more instruction is required.

Notwithstanding the interesting findings obtained from the current study, there are still a few limitations that need to be addressed. The data used in the study were collected over the period of a 15-week semester from a group of native German speakers and another group of upper-intermediate L2 English learners. Whether our findings are generalizable to learners of other languages, of other proficiency levels, and across a longer period of time is yet to be answered. Replication of the current study is thus required. Furthermore, although we were able to answer the question about the interrelationship between accuracy and complexity in L2 learning, we were unable to test the complexity and accuracy trade-off with the L1 German data because of the small sample size. As a result, future research focusing on the interrelationship between complexity and accuracy with more L1 data is on demand. Another limitation of the current study is that the data were from the written mode, making it difficult to investigate the fluency construct, which is the other important component of the CAF triad. Further research is needed with spoken data where the fluency construct can also be accounted for together with complexity and accuracy, hence providing more empirical evidence for theories on the interrelationship among the CAF constructs to guide actual teaching practice.

Chapter 5

Linking Text Readability and Learner Proficiency Using Linguistic Complexity Feature Vector Distance

Chapter highlights

What is already known about this topic:

- Linguistic complexity has been successfully used to predict text readability and assess learner production/proficiency.
- Readability and proficiency are usually represented with one-dimensional label systems like the Common European Framework of Reference for Languages (CEFR) or grade levels.
- Text readability and learner proficiency are both multidimensional constructs like linguistic complexity.

What this study adds:

- We try to link the readability and proficiency spaces with complexity feature vector distances, which keeps the multidimensionality of the constructs and makes the classic SLA theory the Input Hypothesis (or i+1) operationalizable.
- Linking readability and proficiency with complexity feature vector distance makes it possible to provide L2 learners with individualized comprehensible input for proficiency development.

- Validation of the proposed approach provides a solid basis for effective ICALL system design.

Implications for theory, policy, or practice:

- For theory: It makes the Input Hypothesis operationalizable and empirically testable.
- For practice: The study proves the validity of linking the readability and the proficiency spaces with complexity feature vector distances, which forms the basis of an ICALL system design.

Abstract

The automatic analysis of text readability is increasingly tackled with supervised machine learning methods using features derived from the text with natural language processing techniques. Such approaches generally are exclusively based on properties of the texts and they use externally assigned readability levels as gold-standard labels. The readers and the texts readable to them are only indirectly connected through the single, one-dimensional readability label that someone assigned to each text and that was determined to be appropriate for a reader at a certain proficiency level (e.g., determined by a test).

At the same time, texts differ along many different dimensions of linguistic complexity, from morphological, lexical, syntactic, semantic, and discourse aspects of the linguistic system to characteristics of language use—and the language proficiency of readers can also differ with respect to these dimensions. In this study, we therefore propose to link readers and texts directly through multi-dimensional vectors of linguistic complexity measures. We put the idea to a first test by computing the distance between the linguistic complexity vectors for reading texts and texts written by learners. We show that this basic model effectively relates the linguistic complexity development in learner writing and graded readers offering input. The approach makes it possible to empirically investigate the +1 in Krashen's $i+1$, the challenge that best fosters language development given the learner's current interlanguage i. On the practical side, we realize this idea in the ICALL system Syntactic Benchmark (SyB).

We then extend the basic model by linking active and passive language knowledge directly at the level of individual complexity measures. Given the multi-dimensional nature of linguistic complexity on which the input complexity, the learner proficiency, and the +1 can then be determined, it becomes possible to study the impact

of individual (or subsets of) complexity dimensions and to replace the equal weighting of each dimension with a more complex distance measure supporting different degrees of challenges for the different dimensions of linguistic complexity. To illustrate the value of the fine-grained analysis, we analyze the Wang and Wang (2015) continuation writing data and show that substantial alignment between input and output can indeed be observed for most dimensions of linguistic complexity.

Related publication

This chapter is based on the following publication:

- Chen, X. and Meurers, D. (2018a). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, In press.

5.1 Introduction

Target language input at an appropriate difficulty level with regard to the learner's current proficiency level is an indispensable part of language acquisition. The influential Input Hypothesis (IH) of Krashen (1985) emphasized the role of *comprehensible input* at the level of $i+1$, where the i refers to the level of the learner's current interlanguage and the +1 is the next stage of language development. While other authors have highlighted the importance of interaction (Long, 1996), output (Swain, 1985), and noticing (Schmidt, 1995), there is a general consensus that input plays a fundamental role in second language acquisition. Reading texts that are just above the level of the learner's interlanguage has been argued to be particularly important (Rodrigo et al., 2004; Jeon and Day, 2016). Such texts enable the learner to practice being competent readers and motivate them to read more (Milone and Biemiller, 2014), thus exposing them to more comprehensible input exhibiting a broader variety of language and increasingly elaborate forms, hence further promoting their proficiency.

To obtain reading input adapted to the learners' proficiency level, one needs to model the linguistic characteristics of the input, which is related to assessing the readability of the text (Nelson et al., 2012). Going beyond the domain of foreign language learning to academic language development in general, publishers in the US commonly label texts to indicate the grade level of the readers the texts are intended for, and the US Common Core State Standards (CCSSO, 2010) 'call for *a staircase of increasing complexity* so that all students are ready for the demands of college- and career-level reading no later than the end of high school.'[1] Text readability is commonly defined as the sum of all elements of a text that affect a reader's understanding, reading speed, and level of interest in the text (Dale and Chall, 1949). Assessment of text readability can be done qualitatively (Pearson and Hiebert, 2014) or quantitatively (Benjamin, 2012; Collins-Thompson, 2014; Zakaluk and Samuels, 1988), with the latter being considered more objective and easier to automate in order to support the analysis of large numbers of texts.

Readability had traditionally been characterized using regression formulas based on surface features of a text, such as sentence length and lexical difficulty (cf. DuBay, 2006). More recent research makes use of NLP and ML technologies to assess readability automatically (e.g., Vajjala and Meurers, 2012; McNamara et al., 2014; Flor et al., 2013; Williams et al., 1994). NLP technology makes it possible to automatically extract a range of textual features. On this basis, supervised ML then turns

[1] `http://www.corestandards.org/other-resources/key-shifts-in-english-language-arts`

readability assessment into a classification problem in which features of the texts are used to identify the reading levels of the texts. Linking a broad spectrum of features extracted using NLP with ML to combine these observations makes it possible to construct comprehensive models of the linguistic complexity of reading materials taking into accounts features of multiple linguistic levels such as lexis, morphology, syntax, and discourse (Lu, 2010, 2011; Lu and Ai, 2015; Ortega, 2015; Mazgutova and Kormos, 2015; Jarvis, 2013; Kyle and Crossley, 2015). Systems based on these technologies have been found to be effective and accurate in assessing text readability (Nelson et al., 2012; Vajjala and Meurers, 2014).

Strikingly, most readability research exclusively focuses on the properties of the text. Yet, whether a text is comprehensible and at the level of $i+1$ in the sense of the IH, where the i refers to the level of the learner's current interlanguage and the +1 is the next stage of language development, is determined not only by the characteristics of the text itself, but also by the language proficiency of the reader and their previous knowledge of the subject domain providing top-down predictions to be integrated with the bottom-up information from the reading process. Mesmer et al. (2012) proposed to distinguish *text complexity* from *text difficulty*, with the former referring to the lexical, syntactic and discourse features of a text and the latter taking into account the readers' performance on certain tasks based on the text. However, despite Mesmer and her colleagues' theoretical model, there has been little research on how this distinction could be implemented to develop intelligent reading systems capable of assigning reading texts of appropriate difficulty levels based on an individual learner's language proficiency. The readability of a text generally is connected to the proficiency of a reader through an externally assigned label, e.g., a school grad level or a CEFR proficiency level (Council of Europe, 2001), which then is given two interpretations. On the one hand, it represents a proficiency level. On the other, it labels texts for readers who are at that proficiency level.

Reducing the multi-dimensional readability characteristics of a text to a single CEFR or school grade label in this way is problematic since learners with the same proficiency label vary in terms of which aspects of language they are familiar with and can reliably process. For example, second language learners at the CEFR B2 level, who live in an environment where the language is spoken, will clearly exhibit a very different mastery of vocabulary than foreign language learners at the same overall level who never lived in such an environment. On the flip-side of the coin, a text may be challenging for very different reasons, for example, because of the use of complex grammatical structures or due to high lexical diversity. As a result, the one-dimensional nature of the labels of text readability and learner proficiency

makes them suboptimal for the selection of $i+1$ input for individual learners with different proficiency characteristics.

A prominent strand of second language acquisition research characterizes the development of language proficiency in terms of the CAF of the language produced by a learner (cf., Housen et al., 2012; Lambert and Kormos, 2014; Vercellotti, 2017, and references therein). Complexity, which is defined as the elaborateness and variedness of language production (Ellis, 2003), is the most researched construct in the CAF triad. It has been widely used to gauge the learners' proficiency and benchmark the development of their interlanguage (Ortega, 2012). This opens up the possibility of using the linguistic complexity measures developed in SLA research to measure the linguistic complexity of reading material (Vajjala and Meurers, 2012). Taking this perspective one step further, we can use the same measures of linguistic complexity to connect reading materials and readers by comparing the linguistic complexity of the text input with that of texts written by the reader. While the impact of the writing task (Alexopoulou et al., 2017) and the gap between active and passive language proficiency must be taken into account when interpreting learner texts as a multi-faceted record of the reader's proficiency, connecting text and learner in this way seems to be realistic. The active-passive gap can be empirically determined, and the task can be chosen to be comparable across learners and in such a way that it enables the user to showcase what they are capable of (which may well require multiple tasks to cover the full breadth of the elaborateness and variedness of language).

In this study, we explore this idea of linking text readability with learner proficiency by performing a broad range of analyses of the linguistic complexity of both the reading input and the learner production. We use multidimensional vectors[2] encoding the analysis results for the same, rich complexity feature set for analyzing both the input texts and the student writings in order to represent both the readability and the proficiency constructs in terms of comparable dimensions. We show that by calculating the distance between the vectors representing the text readability and the learner proficiency constructs, one can effectively link the two. On the practical side, this approach can be applied to develop ICALL systems for selecting comprehensible reading input that target individual learners. On the foundational

[2]In mathematics, a vector is defined as an object that has both a magnitude and a direction. It can be represented in a coordinate system as a set of coordinates. For example, a two-dimensional vector can be represented as a pair of coordinates (x, y) in a Cartesian coordinate system. The same vector can also be visualized as an arrow pointing from the origin $(0, 0)$ to the point at (x, y) in the Cartesian coordinate system. The *magnitude* of this vector is then the length of the arrow and the *direction* the direction the arrow points to. A vector is not limited to two dimensions; it can be multidimensional, but the basic principles are the same.

side, it makes it possible to empirically investigate the impact of different aspects of input complexity on the language produced by learners.

In what follows, in Section 5.2, we will first present an ICALL approach that takes into account the students' language proficiency as linguistic complexity feature vectors when trying to automatically select reading texts for language learners. In Section 5.3, we then empirically showcase that computing vector distances on the linguistically rich vectors is fully backward compatible to the traditional analysis of readability and proficiency in terms of a single scale (e.g., grade levels). Section 5.4 then introduces the continuation-writing task of Wang and Wang (2015) as a way to experimentally link input texts and learner writing. We show that fine-grained linguistic complexity analyses of input and learner writing can successfully identify the alignment processes between input and readers, potentially providing a very detailed view of individual language development. Overall, we aim to support a precise operationalization and empirical test of Krashen's $i+1$ hypothesis, and to do so in a way that supports concrete practical use of this perspective in an ICALL application.

5.2 An ICALL approach supporting adaptive reading

ICALL systems use NLP technologies to analyze either native language or learner language (Meurers, 2012). Applications of the former include selecting reading materials at appropriate complexity levels (Collins-Thompson, 2014), providing reading materials with enhanced target structures (Meurers et al., 2010; Reynolds et al., 2014), or automatically generating questions for language learning purposes (e.g., Skalban et al., 2012; Chinkina and Meurers, 2017). As for the latter, automatic learner language analysis is applicable in automatic writing evaluation (e.g., Chen and Cheng, 2008; Shermis and Burstein, 2013), detection of production errors (e.g., Rimrott and Heift, 2008), providing benchmark data on frequent mistakes for language educators (e.g., Granger et al., 2007), or automatically providing corrective feedback on learning tasks (e.g., Ai, 2017; Choi, 2016; Heift, 2004; Amaral et al., 2011). Given the goal of our ICALL approach to link reading input and learner proficiency, it will need to include both types of NLP: analysis of the authentic native language serving as input and of the learner language as fine-grained proficiency indicator.

Besides the NLP capability, ICALL system may also take into consideration learner factors such as the learners' understanding of the domain, their learning

strategies, and the acquisitional stages—though only a few ICALL publications have focused on modeling learner factors (Bull, 1994; Amaral and Meurers, 2008; Michaud and McCoy, 2006; Brown, 2002; Chen et al., 2002; Chapelle and Heift, 2009). The fine-grained complexity indicator of the individual learner proficiency that our approach builds on can be seen as a particular kind of learner model that allows us to make concrete what it means to provide appropriate reading material for a given learner. Modeling the learner in terms of the complexity of the interlanguage i they produce forms the foundation on which we can explore Krashen's idea of providing learners with input at the $i+1$ level. Taking a more social perspective, one could also say that our ICALL approach is aimed at providing learners with input that is within their individual Zone of Proximal Development (ZPD, Vygotsky, 1978).

In order to automatically select reading texts that suit the students' language proficiency, an ICALL system needs to assess their proficiency in some ways. As we just motivated, analyzing students' written output can provide an effective, direct way for this. The ICALL system thus first needs to elicit a piece of writing from the student: an article the student wrote recently or a composition written on the spot based on a prompt given by the system. From this writing, the system then automatically extracts the textual features that are also used to analyze text readability. Using the same feature set for both student writing analysis and readability assessment is essential for making the two vector spaces directly comparable.

By calculating the distance between the vector of the learner writing and those of the reading articles provided by the system, the system can select texts that are closest to the learner writing, i.e., the texts that have the shortest distance to the student writing complexity. Of course, this first, simple picture ignores the fact that there is a *gap* between what a learner can understand and what they can produce—the gap between the *passive* and *active* command of a language. If this gap is not taken into account, the system would underestimate the students' proficiency. However, the magnitude of the gap is an issue that is potentially affected by a number of factors, from individual differences such as the learner's motivation, cognitive capacity or proficiency to their background in the content domains of the reading material. Intimately intertwined with the active-passive-gap issue is the parametrization of the +1, the *challenge* beyond the interlanguage level of the learner that the reading text is designed to offer. For the purpose of our adaptive reading application, the combined distance of the *gap* plus the challenge is the crucial parameter needed to select texts for learners. Fortunately, determining this parameter can be approached as an empirical question, which we can answer by varying the parameter in the ICALL system to study which gap+challenge best

fosters learning for different individuals. If we also collect individual difference measures for these individuals, it becomes possible to generalize the collected gap parameter to groups of individuals.

There is an additional granularity issue resulting from the fact that the complexity of a text and the proficiency of a learner can be characterized at each individual complexity feature level (which can be aggregated in different ways, cf. Chen and Meurers, 2017b) or at the combined, text vector level. Both gap and challenge can be computed at either level. At the text vector level, it also is possible to weigh different features differently and, on the more technical side, to compare vectors using different vector distance measures. As a first step, equal weighting of all features seems like a sensible starting point, leaving room for future research providing a more optimized weighting to be learned as another parameter of the ICALL system.

To support the process of second language learning in real life, the provision of adaptive input naturally needs to be ensured over an extended period of time. The overall system setup therefore iterates the writing and reading selection steps as illustrated in Figure 5.1. In section 5.4, we will turn to the special case of continuation writing activities, which actually make it possible to integrate the two steps into one activity and immediately monitor the complexity alignment between input and learner writing as an approximation of learner uptake.

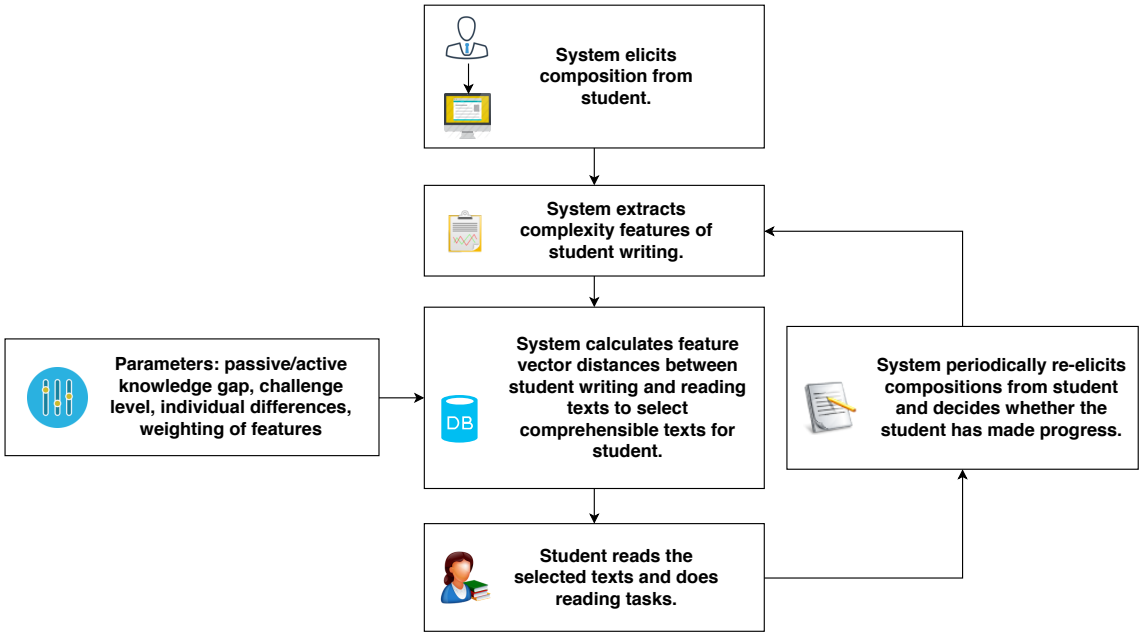

Figure 5.1: An ICALL framework for adaptive reading

We developed a prototypical implementation of the ICALL framework support-

ing adaptive reading called SyB[3], originally focused on individual complexity measures (Chen and Meurers, 2017a, see also, Chapter 6 for detailed introduction). The system provides learners with reading input that matches their proficiency either by analyzing single syntactic complexity features, or by calculating the combined vector-distance between the complexity of the learners' writings and that of the texts from the reading input corpus. Learners first paste in their writing into the system (Figure 6.1), which extracts the complexity features of the writing. SyB then estimates the position of the learner's proficiency level on a scale calculated from the target proficiency of a corpus of leveled reading texts (Figure 6.2). The system then suggests reading material to the learner based on the selected complexity feature or the general text complexity using the vector distance approach proposed in the current study (Figure 6.3).

The system allows learners to manually set the challenge level, realizing a basic, manual parametrization of the combined passive/active knowledge gap and challenge level. The manual selection of challenge level by the learners allows them to obtain texts that they perceive as appropriate challenges, implicitly taking into account individual differences. Ideally, an ICALL system supporting adaptive reading input selection should be able to set these parameters automatically based on its modeling of each individual learner. An empirical study systematically testing the effects of different challenge levels on individual proficiency improvement while taking cognitive individual difference measures into account is reported in Chapter 7.

In the next section, we move from the ICALL system functionality to the concrete level by spelling out how texts are represented by multidimensional feature vectors and how the vector distances are calculated. To evaluate the method, we present an experiment substantiating that the vector distance is a meaningful way to aggregate differences between linguistically fine-grained complexity analyses in a way that makes them interpretable as overall level differences. In section 5.4, we then zoom in on the individual complexity feature levels of input texts and reader productions. The continuation writing task makes it possible to provide a fine-grained analysis of complexity alignment as an indicator of learner uptake, which we evaluate in our second experiment. Taken together, the two experiments help confirm the validity of using linguistic complexity features and text feature vectors as a link between text readability and learner proficiency.

[3]Available at `http://complexityweb.org`

5.3 Linguistic complexity feature vector distance and text readability

5.3.1 Linguistic complexity feature vectors and their distances

One key question that the approach presented in the previous section needs to address is whether the distance between complexity feature vectors of learner-produced text and those of authentic reading material can be meaningfully used to determine which readings are appropriate for the reader. Textual features automatically extracted with NLP tools have been successfully used to assess both text readability (e.g., Crossley et al., 2007; Flor et al., 2013; Lu et al., 2014; François and Watrin, 2011; Hancke et al., 2012; Heilman et al., 2007) and student writings for proficiency placement purposes (e.g., Lu, 2010; Attali and Burstein, 2006). In these studies, the extracted feature values form multidimensional vectors to characterize the readability of reading articles *or* the proficiency of the learner from a writing quality perspective. However, to the best of our knowledge, no attempt has been made to use such feature vectors to directly link the text readability and learner proficiency spaces.

As motivated in the introduction, both the text readability and learner proficiency constructs are multidimensional in nature. Yet, previous research on text readability and learner proficiency generally reduces them to single readability and proficiency label such as Lexile scores (Lexile, 2007) or CEFR labels. This reduction means that we miss out on the opportunity to directly link the learner to the reading materials that best fosters their language development based on empirically observable language properties encoding the broad spectrum of linguistic complexity. We argue for abandoning the reductionist approach aligning proficiency and readability on a single grade or proficiency scale. Instead we keep the rich representation of multidimensional language complexity feature vectors to encode both the text readability and learner proficiency constructs and to use the vector distance between the constructs to relate the two spaces. We first show that our approach remains capable of accounting for readability difference just like the reductionist approach. In section 5.4, we then showcase how the multi-dimensional encoding additionally makes it possible to directly observe the effects of input on production.

5.3.2 Experiment 1: Feature vector distance on a leveled reading corpus

In the first experiment, we tested whether feature vector distance can successfully identify reading level differences between texts for which gold-standard labels are provided using a traditional one-dimensional readability scale. If the distance between textual feature vectors can be used to measure readability level differences, we should be able to show that the vector distances are positively correlated with the level difference between texts in the corpus, i.e., the greater the vector distance in the corpus, the larger the level difference in the corpus, and vice versa.

The traditional approach of using scales such as CEFR to label both learner proficiency and text readability in principle also satisfies the just-mentioned condition if the level labels are considered as interval or ratio variables. However, in addition to some foundational questions about the empirical validity of such scales (Wisniewski, 2017), they also are very difficult to use in practice as part of an ICALL system, given the lack of freely available tests for determining learner proficiency and the only indirect connection to the level of reading material appropriate for learners at a given level. On the methodological side, it is also problematic to consider such scales as interval or ratio variables. For example, one would be hard-pressed to show that the difference between A1 and A2 learners arguably is the same as that between B1 and B2 learners in any clearly quantifiable way. In contrast, both texts written by learners and texts considered as reading material can be represented by multi-dimensional complexity vectors and be straightforwardly compared, both at the level of the individual complexity dimensions and in terms of the overall distance between two text vectors.

The leveled reading corpus In order to test the validity of the proposed method, a leveled authentic texts corpus was used to verify the condition. The validation corpus consists of authentic reading articles targeting learners of different reading abilities from Newsela[4], an educational website that provides reading articles on various topics for language learning. Each article in Newsela is offered in five reading levels marked with Lexile scores. The Lexile score is computed with the Lexile formula of text readability, which is one of the traditional readability formulas constructed by regressing text readability on a few textual features like sentence length and frequency of words from a reference corpus. It is thus an aggregate measure of text readability similar to grade level or age group. The Newsela website also provides a mapping between Lexile scores and US grade levels. For each of the five

[4] `https://newsela.com`

levels, thirty articles were randomly selected from the Newsela website to create a leveled text corpus totaling 150 texts, with an average text length of 763 words.

Complexity features used and vector distance calculation The feature extractors we used are freely available on the Common Text Analysis Platform (CTAP, Chen and Meurers, 2016b), which features a user-friendly web-interface and modularized analysis components for common text analysis needs. We extracted 576 lexical, syntactic, discourse, language use, and traditional surface features (see Appendix A for a partial list of the features) from each text, resulting in a 576-dimensional vector for each text. For each dimension, we encoded the standardized value of a complexity feature, i.e., the z-score encoding how far a given value is from the mean in terms of standard deviations.

While there are multiple options for computing the distance between text vectors, including the possibility of weighting individual dimensions mentioned in the previous section, we settled for the most common Euclidean distance that seems well-suited for the vectors at hand (which are not sparse or have other properties making another measure more appropriate). Each point in the Euclidean n-space is defined by an n-dimensional Euclidean vector. The Euclidean distance d between points p and q in a Euclidean n-space is given by the Pythagorean formula shown in Figure 5.2.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

$$\text{where, } p = \begin{bmatrix} f_1 = p_1 \\ f_2 = p_2 \\ \vdots \\ f_{576} = p_{576} \end{bmatrix} \text{ and } q = \begin{bmatrix} f_1 = q_1 \\ f_2 = q_2 \\ \vdots \\ f_{576} = q_{576} \end{bmatrix}$$

Figure 5.2: Euclidean distance between two vectors p and q representing the linguistic complexity of two texts

We calculated the Euclidean distances between the five levels of the same article, i.e., the complexity vector distances between the texts at levels 1 and 2, at levels 1 and 3, ..., 4 and 5. The proposed condition would be supported if we found that greater level differences (e.g., the distance between levels 1 and 5 as compared to that between levels 1 and 2) are associated with greater Euclidean distances between the text complexity vectors.

Results Figure 5.3 shows a box plot of the textual vector distances with regard to the texts' level differences. Level differences refer to the differences between the

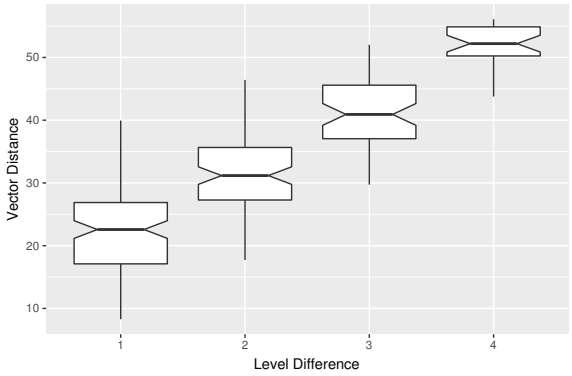

Figure 5.3: Feature vector Euclidean distance on text level difference

levels that the texts target. For example, adjacent levels have a level difference of one while distant levels have higher level differences. The figure shows that the greater the level differences, the further the vector distances.

One-way ANOVA confirmed the differences among vector distances across level differences, $F(3, 296) = 242, p < .001$. Post hoc Tukey HSD tests suggested that significant vector distance differences were found among all level difference pairs (all adjusted $p < .001$). Pearson's correlation coefficient between level difference and vector distance was also highly significant, $\rho = .80 (p < .001)$.

The results confirm that the vector distances between texts represented as multidimensional linguistic complexity vectors are highly correlated with the level differences in a leveled authentic text corpus. It thus is possible to move to a multidimensional representation of text complexity in a way that is fully backward compatible with one-dimensional scales traditionally used to link the readability of texts to the level of learners proficient enough to read those texts.

5.4 Experiment 2: Directly linking learner input and output complexity

Given that in our approach the same representations are used to represent the complexity of (a) texts written by learners, as proxy for their proficiency, i.e., the learner's interlanguage i and of (b) texts to be read by learners as i+1 input fostering the learners' acquisition, we can explore whether using the same representational means for both actually makes it possible to observe a direct, empirical influence between the input learners read and the complexity of their writing. We first inves-

tigate this in terms of the overall text complexity, using vector distance as before. Then we zoom in on the individual complexity measures to test whether they are cognitively real in that input impacts output.

Continuation Writing To explore the impact of learner input on learner output, we need a setup in which learners systematically produce text following exposure to input. Wang and Wang (2015), studying the effect of alignment on L2 production, proposed the CW task that is very well-suited for our purposes.

In the Wang and Wang (2015) study, the CW tasks consist of two English stories from which the endings were removed. Each story was also translated into Chinese. Forty-eight Chinese EFL students, who had learned English for at least seven years but had never been abroad, were asked to read one story in English and the other in Chinese and to write endings for each story in English. The students on average wrote 641 words per text. Wang and Wang show that after reading the beginning of the story in English, participants made significantly fewer errors when writing the end of the story in English than when writing the end of a story in English for which they had read the beginning in Chinese. They conclude that the English reading input provided the students with more target-like language to align their target language production to, hence producing more target-like output, containing fewer errors.

Wang and Wang kindly shared their CW corpus with us for further analysis. We analyzed the English input and writings of the learners in the CW corpus to illustrate that learner proficiency and text readability can be meaningfully linked using vector distance, and to showcase that the detailed complexity analysis can provide a fine-grained perspective of the alignment between input and learner writing.

Given that the English stories used as reading material were chosen by experienced teachers with knowledge of the English proficiency of the students and were intended to foster their English acquisition, we visualize the overall corpus setup assuming that the input texts are at a level above that of the student writing. We then can picture the overall CW corpus with English input texts and the two kinds of English output written by the learners as shown in Figure 5.4.

We consider the student writing after reading the Chinese text as a baseline and the English text input as the highest in complexity. The student writing after reading that English text then should be in-between their baseline writing and the English input text complexity.

If our complexity analysis approach and the above assumptions are correct, the vector distance between the writing after reading the Chinese story and the input

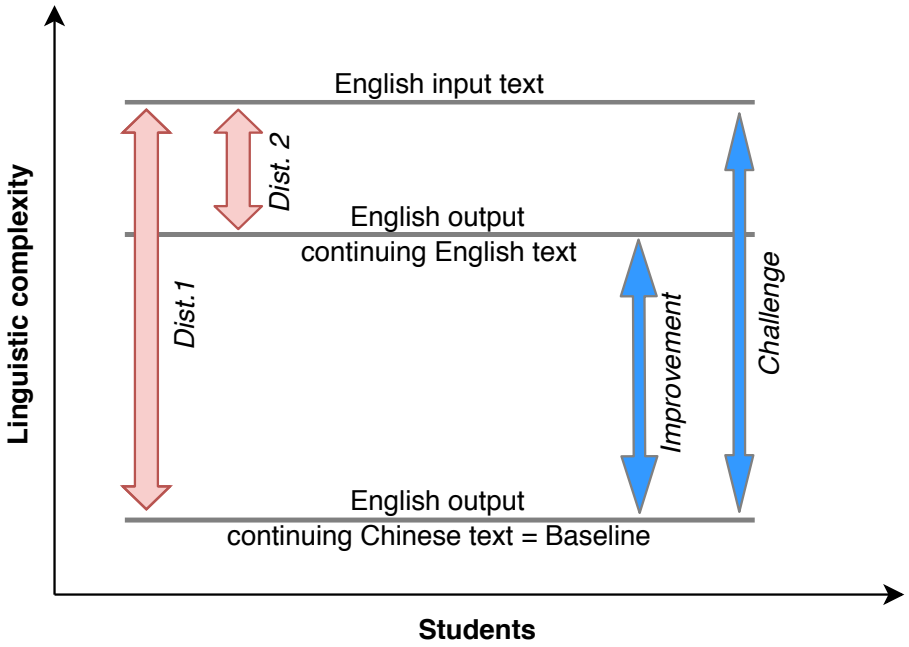

Figure 5.4: Linking complexity of input and output in continuation writing

text (Dist. 1) should be greater than that between the writing after reading the English story and the input text (Dist. 2), i.e., $Dist.\ 1 > Dist\ 2$.

5.4.1 Linking the overall input and output complexity

For each student, we calculated the Euclidean distances between the student's continuation writings under two conditions and the English input text to compute the two distances. Table 5.1 summarizes the results.

	Distance 1	Distance 2
mean	44.59	39.69
sd	6.98	8.21
Paired sample t-test: $t = 5.02, df = 47, p \leq .001$		

Table 5.1: Overall complexity comparison for the CW corpus

The mean distance for Distance 1 was $M_{dist.1} = 44.59(sd = 6.98)$, confirming a gap between the complexity of the baseline student writing and that of the provided English reading material. For Distance 2, between the English input and the student writing continuing that input, the gap was reduced to $M_{dist.2} = 39.69(sd = 8.21)$. A paired-sample t-test confirmed the significance of the difference between the two distances ($t(47) = 5.02, p \leq .001$). The students thus indeed aligned the complexity of their writing with that of the English input.

In terms of the big picture, the analysis confirms that one can meaningfully employ complexity feature vector distances to compare the reading level of reading material and the learners' language proficiency as manifested by the language they produce.

Going beyond the overall alignment visible in the vector distances between the complexity feature vectors of learner input and output, we can now investigate in which of the dimensions alignment can be observed. In other words, we can empirically determine, which of the many different aspects of linguistic complexity of the input are cognitively real in that they were (implicitly) perceived by the learner, allowing them to adjust their writing accordingly.

5.4.2 For which individual dimensions of linguistic complexity can alignment be observed?

The development of learner proficiency is manifested in their mastery of a wider range of linguistic structures or more elaborate ways to express ideas when comprehending and producing the language (Ellis, 2003). While we illustrated in the previous section that the overall input and output complexity can be related by computing the distance between the complexity feature vector representations of the texts, a conceptually more relevant question for researchers that the rich, multidimensional representation of complexity allows us to ask is how the complexity of the input at the various levels being modeled affects the production and potentially the development of the learner proficiency. In other words, which of the aspects of complexity is cognitively real in the sense that it can be perceived in the input and the output can be adjusted accordingly.

To pursue this question, we analyze the CW corpus data in terms of two indices, *challenge* and *improvement*, as illustrated by the blue arrows in Figure 5.4 and spelled out in Figure 5.5.

$$\begin{aligned} \text{challenge} &= \text{complexity(English input)} - \text{complexity(baseline output)} \\ \text{improvement} &= \text{complexity(continuation output)} - \text{complexity(baseline output)} \end{aligned}$$

Figure 5.5: Defining *challenge* and *improvement* in terms of input and output of learners in CW corpus

For a given student, the *baseline output* is the student writing after reading the Chinese input. The *challenge* the student faced is the difference between the complexity of the English reading input and that of their baseline writing. The *improvement* is the change in complexity of their writing after reading the input text compared to that of their baseline writing. We want to find out for which

complexity features the nature of the *challenge* predicts the *improvement* in the learner production.

A linear regression model was fitted by regressing improvement on challenge for each individual textual feature. We found that for the majority of features (403 out of 576, i.e., 70% of all features, see Appendix C for detailed statistics of the fitted models), the challenge was able to explain variance in the improvement, with the model fit measure of R-squared ranging from 0.94 to 0.08. In other words, the *challenge* was highly predictive for the *improvement*—up to 94% of the variance in the *improvement* was predicted by the *challenge*. The estimated statistics for the slope (β) were all positive, so the more challenge a learner received from the input with respect to the complexity of the student's writing, the more complex their writing is going to be after reading the challenging input.

Our results resonate with the learner error analysis results of Wang and Wang (2015), who explain the phenomenon as an alignment effect. Wang and Wang's analysis based on the number of errors made by the students only illustrated positive alignment, though, i.e., students made fewer errors after reading the English text than after reading the Chinese text. In our analysis of the different facets of complexity, we observe alignment throughout the spectrum of challenges in complexity, both positive and negative. That is, when the complexity of the input is the same as or below what the students are already capable of producing—thus offering no challenge or a negative challenge—the complexity of their continuation writing also goes below their baseline production, resulting in negative improvement of complexity.

In Figure 5.6, we illustrate the close relation between input challenge and improvement of student writing for four complexity measures drawn from the result table in Appendix C. The top left panel shows the most correlated feature, the number of verb types in their third person singular present forms, for which 94% of the variance in the improvement is explained by the challenge. Third person singular present inflection of verbs is a challenge to Chinese students of English because Chinese is a non-inflectional language. However, verb inflections are usually explicitly taught at an early stage of the school curriculum. So it is likely they are consciously aware of the need to pay attention to this language aspect. At the same time, it is not the case that all readily observable language characteristics show alignment. On the top-right, we see the global noun overlap—a measure of textual cohesion that is readily observable, yet it does not even show a significant correlation, with only 1% of explained variance.

As an illustration of a complexity measure at the lexical level, consider lexical

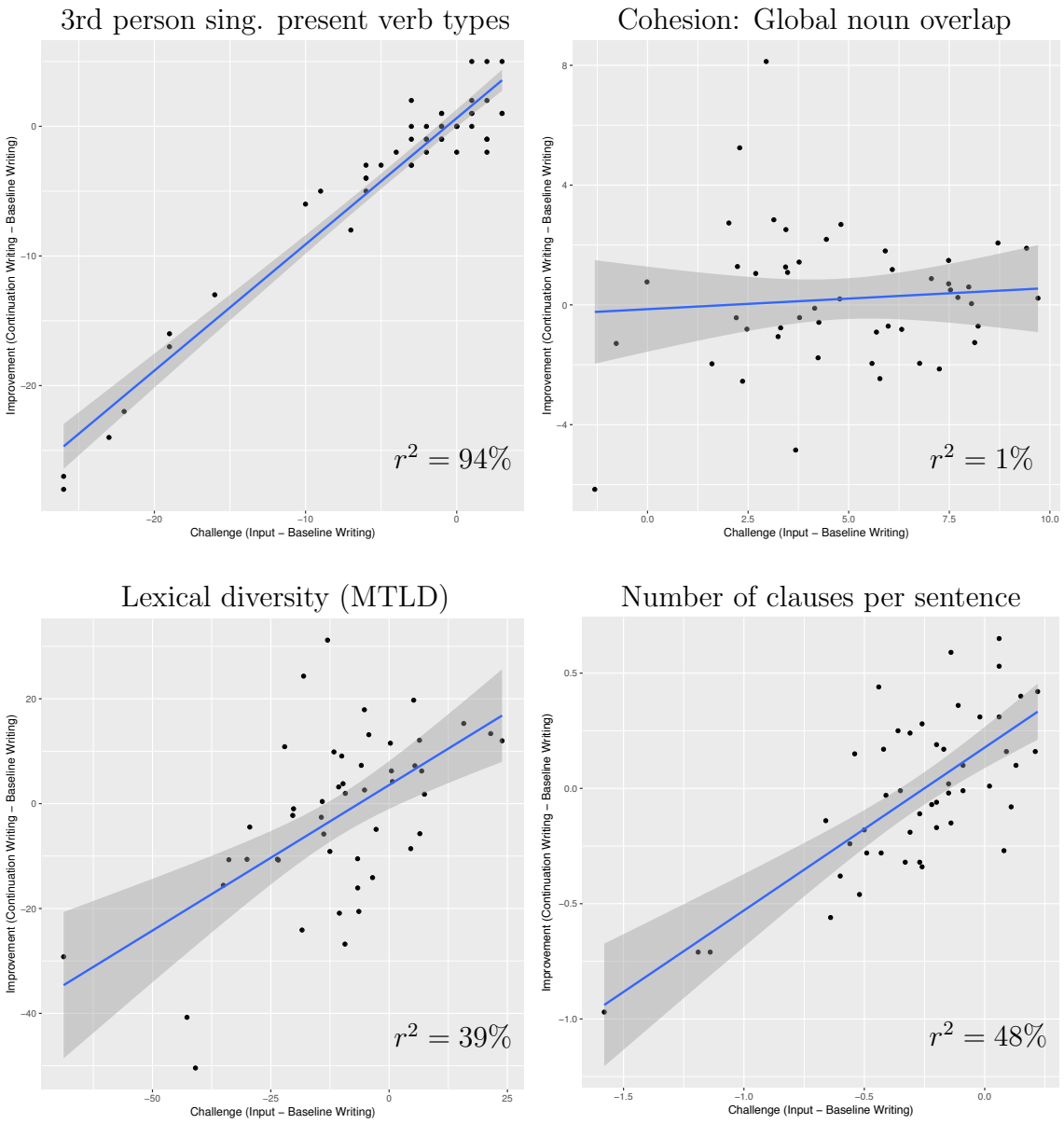

Figure 5.6: Relationship between improvement and challenge for four linguistic complexity features

diversity as measured by MTLD (McCarthy and Jarvis, 2010). It shows a clear positive correlation with an R^2 of 39%. Interestingly, the lexical diversity of the input text for most learners was not challenging, though, with most learners being underchallenged. So here the active language use by the learner in the baseline condition actually was more lexically diverse than the reading material given as input.

An example for syntactic complexity is the mean number of clauses per sentence shown at the bottom right, which shows a higher amount of explained variance (48%). Again, the input material only offered a challenge to very few of the learners, so that most of the correlation is due to alignment of the writing down to the complexity of the input. While it illustrates the ability of the learner to align the complexity of their writing to be appropriate for the continuation writing task, incrementally more challenging input would be needed to show alignment and learning gains building on each other step by step.

The data in the CW corpus was collected with only one challenging input for each student, so that it is not possible to see if different challenge levels would result in different levels of improvement for a given, individual students. We would expect that the improvement line will not keep increasing towards the positive end of challenge. If the learners were challenged to a degree that is beyond their 'Zone of Proximal Development', in the sense that the more elaborate or varied ways of writing something down have not yet been acquired or the overall, combined complexity becomes too high to handle for the learner, the improvement would be expected to level off. An intervention study to investigate the effects of different levels of challenge on complexity development is currently being conducted to investigate this further.

Summing up the second experiment, our analysis of the CW data shows that the learner input and output spaces are clearly linkable with the textual complexity features at very fine-grained levels. The approach therefore can be seen as offering an empirically grounded, effective operationalization of the concept of $i+1$ input that can be parameterized and tested with a broad range of linguistic complexity factors.

5.5 Summary

Language learners need authentic target language input that matches their language ability and interests. Readability analysis can be carried out to assign texts to readers, but traditionally the outcome of such analysis is a one-dimensional label of

the grade or proficiency level that a text is deemed to be suitable for. Yet, second language learning is well-known to be highly variable, with individual differences playing an important role—and linguistic complexity is known to be a highly multidimensional construct involving all aspects of the linguistic system, language use, and human sentence processing. Using a single, one-dimensional label to link readers to their reading material therefore is inadequate.

The present study approaches the problem by taking the multidimensionality of both the readability and the proficiency constructs into account. We show that multidimensional complexity feature vectors can be used to represent the two and vector distance readily supports relating them at an aggregate level. The approach was validated using an authentic reading corpus targeting different learner levels, and we showed that the same method can be used to link learner input and output complexity based on the data from a continuation writing corpus. Excitingly, the learner and the reading input then can also be related at the fine-grained individual linguistic complexity feature level. On the practical side, the proposed approach provides the foundation needed to develop an ICALL approach for reading text selection based on learner proficiency, for which we worked out the architecture and illustrated it with the SyB system prototype (see the next Chapter).

While we have focused on selecting reading material for learners based on their language proficiency, teachers will also want to pursue a pedagogical focus in their courses and select materials on certain aspects of the target language at a given time. For example, when focusing on vocabulary learning one may want to select texts that repeat the target words multiple times to increase exposure to the target vocabulary (Cobb, 2007; Ghadirian, 2002). Language-aware search engines such as Form-Focused Linguistically Aware Information Retrieval (FLAIR) (Chinkina and Meurers, 2016) can support teachers in ensuring a rich representation of the pedagogically targeted language constructions. Fortunately such a pedagogical input enrichment approach reranking search results is fully compatible with the SyB approach selecting reading texts based on the syntactic complexity of the learner production and the intended challenge level. Indeed, any comprehensive approach will need to integrate all three: an analysis of the reading material, of the learner characteristics, and of the pedagogical agenda.

The contribution of this study to language learning research is that it provides a way to relate learning input to learner production, making it possible to empirically investigate Krashen's (1985) $i+1$ hypothesis. The broad complexity feature representation of learner output provides a rich and fine-grained characterization of the learner's interlanguage i. It also makes it possible to characterize potential reading

input for this learner using the same dimensions of linguistic complexity so that it can be directly related to the characterization of the learner. In future research, we can therefore study the effect of different levels of challenge on the complexification of the learner's language use and potentially the development of their interlanguage. Given the framework laid out in this study, the optimal individual parametrization of the challenge, how much the +1 should be for a given dimension of complexity, has turned into a question that can be empirically studied now. The fact that alignment between learner input and subsequent learner production was readily apparent for most complexity features is very promising—though it remains to be seen, whether such alignment also is incrementally increasing through longitudinal exposure and whether it leads to language learning in the sense of increasing the learner's ability to complex language as appropriate for a given task.

The current study also contributes to the field of Computer Assisted Language Learning (CALL) by demonstrating how NLP methods can be utilized to address real-life, conceptually grounded challenges in language education. We combined findings in second language acquisition research with research on text readability and learner writing assessment with NLP methods to develop an ICALL solution for automatic selection of reading texts. A prototype instantiating the approach has been developed with the SyB system (see the next chapter), and experiments investigating its effectiveness have also been conducted and reported in Chapter 7. We are confident that the multi-disciplinary grounding and integration of perspectives of the presented approach provides a solid foundation for the further development and use of ICALL addressing established needs in second language teaching and learning.

Chapter 6

SyB—An ICALL System for Developing Syntactic Complexity

Chapter highlights

What is already known about this topic:

- Successful SLA depends a lot on the challenging input a learner receives. This input is often denoted as $i+1$ or input that is within the learner's Zone of Proximal Development. Both the linguistic complexity of the text and the developmental stage of the learner decide the challenge levels of the input.
- Syntactic complexity of reading input and learning production is an important measure for determining the appropriateness of the input challenge and the developmental level of the learner. Automatic system for the analysis of syntactic complexity of learning input and learner production has been successfully developed.
- As has been shown in the previous chapter, the input and proficiency spaces are linkable with the multidimensional complexity construct.

What this study adds:

- An ICALL system that automatically assigns reading input to learners based on the system's assessment of the syntactic complexity of the learners' production was developed.
- The system uses a pedagogic corpus, instead of the common practice of using learner corpus, to create the developmental benchmark of syntactic complexity to account for the learners' developmental stages.

- The ICALL system provides learners with controls over the syntactic challenge levels and the general reading levels of the input texts.

Implications for theory, policy, or practice:

- For theory: The study provides an operationalizable implementation of the Input Hypothesis (i+1), which requires account into both the learner proficiency factor and the textual complexity factor for the selection of comprehensible input for language acquisition purposes.
- For practice: Our system shows how SLA theory and NLP technologies can be combined to develop practical ICALL systems with solid theoretical basis.

This chapter is based on the following publication:

- Chen, X. and Meurers, D. (2017a). Challenging learners in their individual Zone of Proximal Development using pedagogic developmental benchmarks of syntactic complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa*, pages 8–17, Gothenburg, Sweden, 22nd May. Linköping University Electronic Press, Linköpingsuniversitet.

6.1 Introduction

The analysis of linguistic complexity is a prominent endeavor in SLA where NLP technologies are increasingly applied in a way broadening the empirical foundation. Automatic complexity analysis tools such as CohMetrix (McNamara et al., 2014), the L2 Syntactic Complexity Analyzer (Lu, 2010), and the Common Text Analysis Platform (Chen and Meurers, 2016b) support studies analyzing interlanguage development (Lu, 2011; Lu and Ai, 2015; Mazgutova and Kormos, 2015), performance evaluation (Yang et al., 2015; Taguchi et al., 2013), and readability assessment (Vajjala and Meurers, 2012; Nelson et al., 2012).

In this study, we introduce a new system called *Syntactic Benchmark (SyB)* that utilizes NLP to create syntactic complexity benchmarks and identify reading material individually challenging learners, essentially instantiating the next stage of acquisition as captured by Krashen's concept of $i+1$ (Krashen, 1985) or relatedly, but emphasizing the social perspective, Vygotsky's Zone of Proximal Development (ZPD, Vygotsky, 1978).

In terms of structure of the study, we first locate our approach in terms of the CAF framework in SLA research. Then we review approaches adopted by earlier studies in developmental complexity research, including problems they pose for a pedagogical approach aimed at offering developmental benchmarks. We propose and justify a solution, before presenting the architecture and functionality of the SyB system.

6.2 Development of syntactic complexity

The three-part model of development distinguishing CAF has gained significant popularity among SLA researchers (Wolfe-Quintero et al., 1998; Skehan, 2009; Housen et al., 2009; Bulté and Housen, 2012) since it was first delineated by Skehan (1989). It provides SLA researchers with a systematic and quantitative approach to development. Among the CAF triplet, complexity arguably is the most researched and most 'complex' due to its polysemous and multidimensional nature (Bulté and Housen, 2012; Vyatkina et al., 2015). Complexity in the SLA literature has been used to refer to task, cognitive, or linguistic complexity (Housen et al., 2009). In the present study, we investigate complexity from a linguistic perspective, where it is concisely characterized by Ellis (2003) as 'the extent to which language produced in performing a task is elaborate and varied'. While the linguistic complexity construct consists of a range of sub-constructs at all levels of linguistic modeling, such as lexical, morphological, syntactic, semantic, pragmatic and discourse (Lu, 2010,

2011; Lu and Ai, 2015; Ortega, 2015; Mazgutova and Kormos, 2015; Jarvis, 2013; Kyle and Crossley, 2015), the focus in this study is on syntactic complexity.

In line with Ellis's (2003) definition of linguistic complexity, Ortega (2003) characterized syntactic complexity as the range of syntactic structures and the elaborateness or degree of sophistication of those structures in the language production, which we adopt as the operational definition in this study. The uses of syntactic complexity analysis in SLA research include (i) gauging proficiency, (ii) assessing production quality, and (iii) benchmarking development (Ortega, 2012; Lu and Ai, 2015).

The development of syntactic complexity in language produced by learners is closely related to the learner's proficiency development. While the goal of language acquisition is not as such to produce complex language, advanced learners usually demonstrate the ability to understand and produce more complex language. With increasing proficiency, the learners are expanding their syntactic repertoire and capacity to use a wider range of linguistic resources offered by the given grammar (Ortega, 2015), thus producing 'progressively more elaborate language' and 'greater variety of syntactic patterning', constituting development in syntactic complexity (Foster and Skehan, 1996). As a result, syntactic complexity is often used to determine proficiency or assess performance in the target language (Larsen-Freeman, 1978; Ortega, 2003, 2012; Vyatkina et al., 2015; Wolfe-Quintero et al., 1998; Lu, 2011; Taguchi et al., 2013; Yang et al., 2015; Sotillo, 2000).

Besides the practical side of performance assessment and placement, in SLA research the developmental perspective is considered to be 'at the core of the phenomenon of L2 syntactic complexity' (Ortega, 2015). However, it is also the least addressed and understood phenomenon of syntactic complexity in SLA research (Vyatkina et al., 2015; Ortega, 2012). Understanding the development of syntactic complexity would enable SLA researchers to determine trajectories of the learners' development and set benchmarks for certain time points or across a given time span. On the practical side, such work could help language teachers select or design appropriate learning materials, and it can provide a reference frame for testing the effectiveness of instructional interventions. Hence researching syntactic complexity from a developmental perspective is of far-reaching relevance and applicability.

6.2.1 Development of syntactic complexity in learner corpora

A number of longitudinal and cross-sectional studies have been conducted to investigate the relationship between syntactic complexity and learner proficiency, aimed at

finding (i) the most informative complexity measures across proficiency levels (Lu, 2011; Ferris, 1994; Ishikawa, 1995), (ii) the patterns of development for different syntactic measures (Bardovi-Harlig and Bofman, 1989; Henry, 1996; Larsen-Freeman, 1978; Lu, 2011), or (iii) discovering a developmental trajectory of syntactic complexity from the learner production (Ortega, 2000, 2003; Vyatkina, 2013; Vyatkina et al., 2015).

With a few exceptions (Vyatkina, 2013; Tono, 2004), one thing these studies have in common is that they analyze the syntactic complexity development of learners based on their production. This seems natural since it investigates complexity development by analyzing the production of the developing entity, i.e., the learners. In principle, a longitudinal learner corpus with a continuous record of productions from individual learners over time would seem to enable us to determine the developmental trajectory and linguistic complexity benchmarks. However, this approach encounters some challenges that make it suboptimal for determining developmental benchmarks in practice.

First, the approach is dependent on learner corpora varying significantly on a number of parameters such as the learners' background, the tasks eliciting the production, and the instructional settings, etc. Significant effects of such factors on the syntactic complexity of learner writing have been identified in a number of studies (Ellis and Yuan, 2004; Lu, 2011; Ortega, 2003; Sotillo, 2000; Way et al., 2000; Yang et al., 2015; Alexopoulou et al., 2017). Consequently, the developmental patterns or benchmarks constructed from different learner corpora, elicited using different tasks, etc. are likely to vary or even contradict each other. For example, the correlation between subordination frequency and proficiency level have been found to be positive (Aarts and Granger, 1998; Granger and Rayson, 1998; Grant and Ginther, 2000), negative (Lu, 2011; Reid, 1992), or uncorrelated (Ferris, 1994; Kormos, 2011). It is difficult to build on such conflicting findings in practice.

Second, the NLP tools used for the automatic complexity analysis do not work equally well when applied to the language produced by learners at varied proficiency levels. Complexity analysis is currently performed using tools developed for different analysis needs (McNamara et al., 2014; Lu, 2010; Kyle and Crossley, 2015; Chen and Meurers, 2016b). They enable fast and robust analysis of large corpora, in principle making the conclusions drawn from these analyses more powerful. However, analyzing learner data can pose significant challenges to the NLP components, which were usually developed for and tested on edited native language, as found in newspapers. While some NLP tools were shown to be quite reliable for analyzing the writing of learners at upper intermediate proficiency or higher (Lu, 2010, 2011),

their robustness for lower-level writing or for some types of task (e.g., not providing reliable sentence delimiting punctuation) is questionable, requiring dedicated normalization steps and conceptual considerations (Meurers and Dickinson, 2017). This may well be why developmental profiling has rarely been done for learner language below upper-intermediate proficiency levels, as Ortega and Sinicrope (2008) observed. This currently limits the possibility of determining developmental benchmarks or trajectories across the full range of proficiency levels.

Last but not least, second language proficiency development is systematically affected by individual differences, making complexity research findings from learner data chaotic and hard to generalize. For example, Vyatkina et al. (2015) observed a 'non-linear waxing and waning' (p. 28) for different modifier categories in a longitudinal learner corpus. Norrby and Håkansson (2007) identified four different types of morphosyntactic complexity development in a corpus of Swedish adult learner language, referred to as 'the Careful', 'the Thorough', 'the Risk-taker', and 'the Recycler'. The analysis of morphological development in English L2 acquisition presented by Murakami (2013, 2016) also highlights the importance of accounting for individual variation in modeling L2 development. As a result, given the current state of affairs and without complex models integrating a range of factors, developmental benchmarks based on learner corpora are of limited practical use for proficiency placement or performance assessment. Naturally this does not mean that research into developmental patterns based on learner corpora is not important or relevant for SLA. On the contrary, the dynamic and adaptive nature of language acquisition means that it is challenging and interesting to approach language development in a way accounting for individual differences (Larsen-Freeman, 2006; Verspoor et al., 2008, 2012), task effects (Alexopoulou et al., 2017), and other factors. For benchmarking and developmental tool development it is useful to look for a more stable data source though.

6.2.2 Developmental benchmarks of complexity in a pedagogic corpus

Considering the challenges just discussed, we explore the analysis of syntactic complexity in pedagogic language corpora compiled from well-edited Target Language (TL). A pedagogic TL corpus is a corpus 'consisting of all the language a learner has been exposed to' (Hunston, 2002), or more realistically 'a large enough and representative sample of the language, spoken and written, a learner has been or is likely to be exposed to via teaching material, either in the classroom or during self-study activities' (Meunier and Gouverneur, 2009). An optimal TL corpus for

benchmarking syntactic complexity development would be one that includes texts targeting learners at any proficiency level, i.e., covering the full spectrum.

The advantages of a pedagogic corpus for developmental benchmarking are two-fold: First, pedagogic corpora can be constructed to exhibit a linear development of complexity measures, as shown by Vyatkina (2013) and confirmed here later. While the developmental trajectory in learner productions is 'bumpy' and influenced by individual differences, task, and other factors discussed earlier, the pedagogic corpus can be written in a way targeting increased linguistic complexity. This is desirable if one wants the class to follow an instructional progression enriching grammatical forms in line with the pedagogic input they receive (Vyatkina, 2013). Pedagogically, it should be easier for language teachers to select instructional materials based on a linear benchmark of linguistic complexity, especially if one has evidence of the students' proficiency using that same scale.

Second, the problem of the NLP tools being challenged by learner language, especially that of the low-proficiency learners, is avoided since pedagogic corpora contain texts with grammatically well-formed and edited articles. Considering the high accuracy of current NLP for such text material, the developmental benchmark constructed from a pedagogic corpus using automatic complexity analysis tools should be highly reliable. It should be acknowledged that no benchmarking system can avoid analyzing learner language if the system is used for proficiency placement purposes (unless additional, external language tests are used). However, complexity benchmarks constructed based on a TL corpus are more reliable than a comparison with a benchmark computed based on learner corpora. If the NLP tools fail to process the learner production to be compared to the benchmark because of grammar errors, resulting in placing the student on a lower level of the TL benchmark, the placement in a sense still is indicative of the aspect of the learner language that needs to be improved.

In sum, the above review suggests that a developmental perspective to syntactic complexity aimed at teaching practice can be meaningfully approached with the assistance of a pedagogic corpus consisting of texts targeting learners in a wide spectrum of language proficiency. In the following section, we will introduce an NLP-based system based on this idea.

6.3 The SyB system

SyB is an ICALL system that analyzes the syntactic complexity of a text produced by a learner and places the text onto a developmental scale constructed from a

comprehensive pedagogic corpus. The system aims at helping learners place the syntactic complexity level of their writings with regard to the pedagogic benchmark and identify the syntactic areas where further improvement is needed. The system is able to visualize the developmental benchmark for different syntactic complexity measures and the learner's position on the benchmark for the selected complexity index. Based on the complexity level of the user's language output, SyB then proposes appropriately challenging texts from the pedagogic corpus. Reading these texts providing i+1 input should help the user advance in language proficiency. The size of the +1, i.e., the degree of the challenge and the overall proficiency level that the learner assumes being at currently are manually specified by the user.

Figure 6.1 shows the Data Window, into which the learner enters a text they

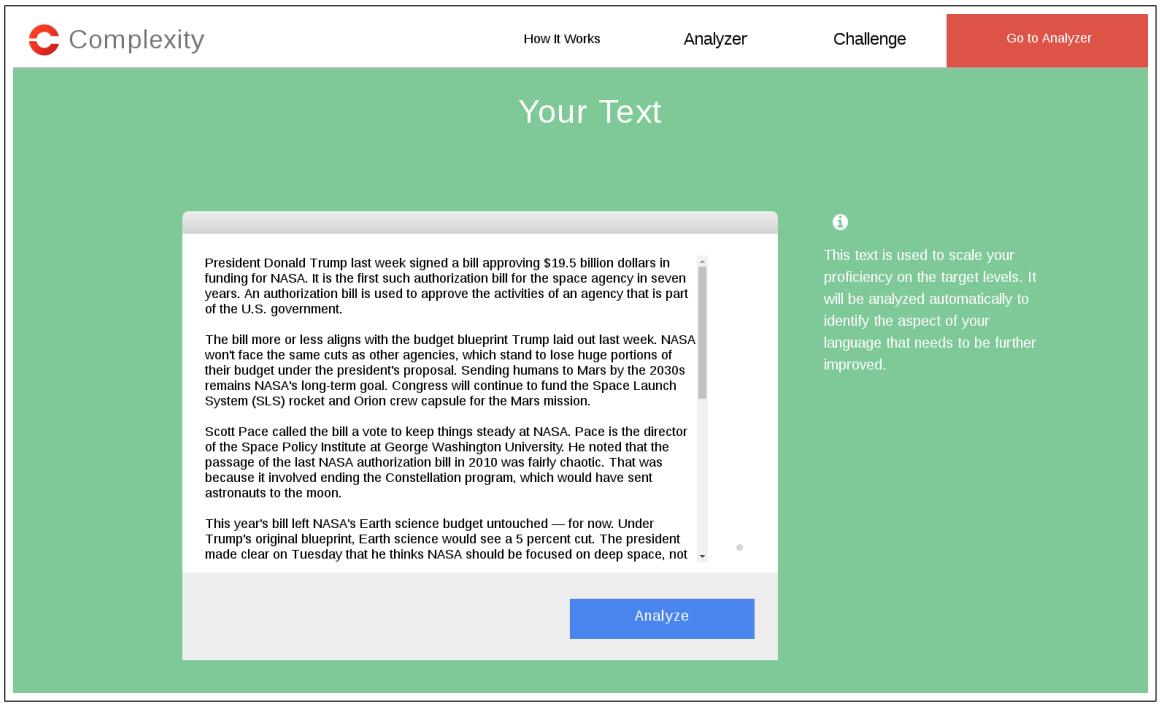

Figure 6.1: The Text Input Window of the Syntactic Benchmark Analyzer, where users can paste a composition to identify their level in relation to the TL benchmark corpus

wrote to identify its level in terms of syntactic complexity in relation to the TL benchmark corpus. In Figure 6.2, we see the Visualization Window providing the result of the analysis for the selected complexity feature (here, the Mean Length of Clause measure). The boxplots show the results for each text in each level in the TL benchmark corpus, and a red line indicates the measure's value for the learner text. Selecting the 'Challenge' button leads to the Search Result Window shown in Figure 6.3. It provides a search result list with links to TL articles intended as $i+1$ input material for the learner. The texts are slightly above the level of

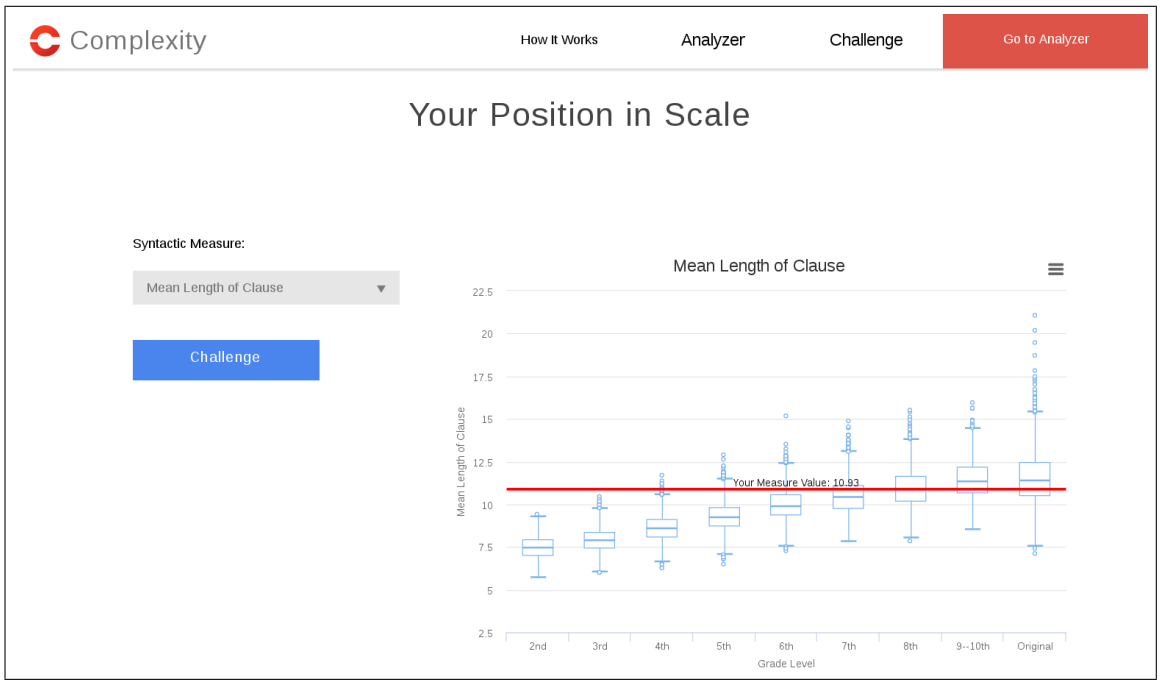

Figure 6.2: The Visualization Window showing the users' level (red line) for the selected syntactic complexity measure (here: Mean Length of Clause) in relation to the TL benchmark corpus

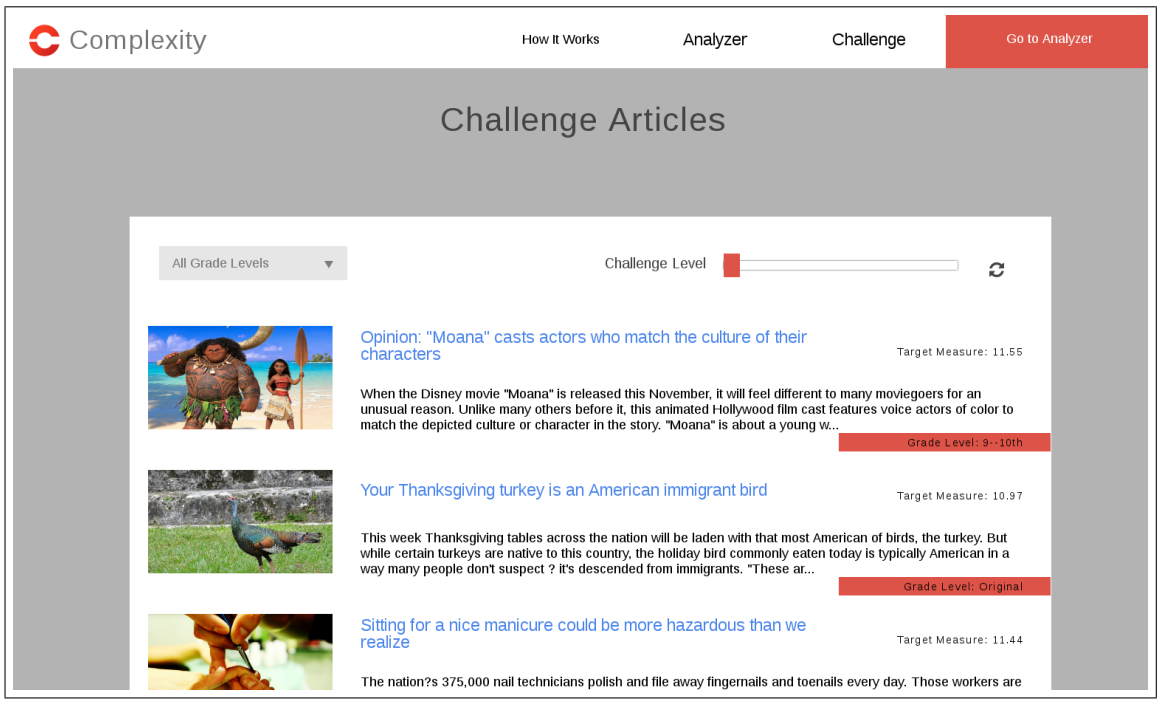

Figure 6.3: The Challenge Window supporting selection of TL articles based on the learner production's syntactic complexity level (and user-specified degree of challenge and overall target grade level)

the learner text in terms of the selected complexity measure, with the degree of the challenge being determined by the user setting. The learner also specifies the overall proficiency level they assume to be in so that the text challenging them in terms of the selected complexity measure is selected from the pool of texts intended for that overall proficiency level.

In the following, we take a closer look at the SyB components.

6.3.1 The pedagogic corpus

The pedagogic TL corpus used for constructing the syntactic complexity benchmark consists of 14,581 news articles from the educational website Newsela[1], which is a website that provides news articles on a wide range of topics. Each article on the website is adapted into five reading levels (including an 'original' level, which is the article in its unadapted form) by human editors. Newsela uses the Lexile Framework (Lexile, 2007) for text leveling and provides a grade to Lexile mapping for converting from Lexile scores to US grade levels. Since the grade level is easier to understand for most users, the SyB system uses grade levels as benchmarking levels. For copyright reasons, the SyB system does not store the original articles from Newsela. It only keeps records of the complexity statistics of the articles and the Search Result Window provides the results in terms of links to the text on the Newsela web site.

6.3.2 NLP processing

Each article in the Newsela TL reading corpus was processed with an NLP pipeline consisting of a sentence segmenter, a tokenizer and a parser from the Stanford CoreNLP Toolkit library (Manning et al., 2014). Tregex (Levy and Andrew, 2006), a utility for tree pattern matching, was used to extract syntactic units such as coordinate phrases, clauses, and T-units from the parse tree of a sentence.

We used the Tregex patterns of Lu's (2010) L2 Syntactic Complexity Analyzer and calculated the same set of 14 syntactic indices suggested in his study (p. 479, Table 1). This set of syntactic features have also been used in developmental syntactic complexity studies and proved to be valid and reliable (Larsen-Freeman, 1978; Ortega, 2003; Wolfe-Quintero et al., 1998). The SyB system currently uses a replication of Lu's processing pipeline, which was shown to have achieved a very high level of reliability in a number of studies (Lu, 2010; Lu and Ai, 2015; Yang et al., 2015; Ai and Lu, 2013; Lu, 2011).

[1] `https://newsela.com`

In future work, we plan to integrate the broad range of linguistic complexity measures offered by our Common Text Analysis Platform (Chen and Meurers, 2016b).

6.3.3 Benchmarking and challenging

For each of the 14 syntactic measures, a benchmark box plot of the measure values by grade level was created. Whenever the user pastes or enters a representative production and chooses the measure they are interested in, the SyB system calculates the chosen measure value from the user text and draws a horizontal red line across the benchmark box plot to signify the relative position of the user text's complexity level on the TL corpus benchmark. Figure 6.2 shows an example of a benchmark plot and the learner text as measured by the same complexity index, Mean Length of Clause.

The system then selects from the TL corpus those articles that challenge the user in terms of specific syntactic complexity as measured by the user's choice of complexity indicator. The user is also given choices of the overall target grade levels of the texts and the level of challenge they want to receive (Figure 6.3). The range of challenge levels matches the range of the syntactic measure calculated from the TL corpus. The complete challenge range is divided into ten sections and controlled by a range slider with those steps, shown as the red slider in the top-right corner of Figure 6.3.

Each article in the Newsela TL reading corpus comes with the overall evaluation of reading level by the editors. Since there is significant overlap in the range of complexity measure values across target reading levels, it is useful to let the user determine the overall pool of texts that they want the system to select from using the selected complexity measure. In SyB, the overall reading level of the challenge texts is selected using the drop-down listbox in the top-left corner of Figure 6.3. The current system then only evaluates a single complexity feature of the learner's production (in the case of Figure 6.2, Mean Length of Clauses) and proposes texts at an appropriately challenging levels based on this single aspect, selected from the pool of texts at the user-selected overall level.

This is not optimal because whether a text poses challenges to specific readers also depend on other factors, such as the lexical complexity, the learners' language competence including aspects such as strategic competence, their world and domain knowledge, and so forth. An alternative method we intend to explore in the future is to compute a broad range of complexity measures using the NLP from our Common Text Analysis Platform (Chen and Meurers, 2016b) so that each text is represented by a vector encoding the results for each complexity measure for that text (which

could also include dimensions for other factors to be considered, such as measures of the user's domain knowledge for different topics or subject domains). The overall $i+1$ challenge can then be computed using a vector distance metric (Manhattan, Euclidean, etc.). Perhaps most attractively, one could combine the two approaches, with the vector-based overall comparison replacing the current manual setting of the global level determining the set of texts to be considered, and the challenge being determined by the user-selected single complexity measure as in the current approach.

The hypothesis behind the overall setup is that by reading the challenging texts, the users will 'align' (Wang and Wang, 2015) to the target levels of syntactic complexity, hence promoting their TL proficiency. Whether this hypothesis is correct and which approach works best for determining input material appropriately challenging learners is an empirical question. Answering it should also provide important insights into the question how Krashen's notion of an $i+1$ (or Vygotsky's ZPD) can be operationalized in terms of measurable features such as linguistic complexity.

6.4 Summary

This study introduced the ICALL system SyB for benchmarking syntactic complexity development based on a TL corpus. A TL corpus can provide a consistent, linear, and complete instantiation of incremental complexification for different aspects of linguistic complexity. Current NLP technologies are more robust for analyzing such TL corpora than for analyzing learner corpora. As a result, syntactic complexity benchmarks in TL corpora may be more applicable and relevant for instructional use than models of linguistic complexification based on learner corpora, which are harder to analyze automatically, exhibit significant individual variation, task effects, and other uncontrolled factors.

However, this hypothesis remains to be validated empirically in actual teaching practice. Future research also needs to investigate which level of challenge for which of the complexity measures at which domain of linguistic modeling is most effective at fostering learning, i.e., what constitutes the best +1 for which aspect of linguistic complexity (for learners with which individual characteristics). Last but not least, while the SyB system provides users with options to control the syntactic complexity and overall reading challenge levels, the system does not take into account the gap between the active ability exhibited in production and the passive ability used for comprehension. The receptive and productive knowledge were found to differ within learners in a number of studies (Zhong, 2016; Schmitt and Redwood, 2011).

It will also be interesting to compare this kind of individual adaptation of the complexity of the input based on the complexity analysis of the learner's production with the input enrichment supported by a teacher-based selection of the constructions targeted to be learned as supported by the FLAIR system (Chinkina and Meurers, 2016).

Finally, it will be interesting to enhance the system by making the texts it suggests for reading adaptive not only to what the learner is capable of producing, but also to how well the learner understands the articles suggested by the system. A production task called Complex Input Primed Writing (cf. Chapter 7) is well suited for this purpose. In a CIPW task the learner is asked to continue writing a text whose ending has been removed after reading the complexity challenge texts. This will make it possible for the system to analyze (i) whether there is uptake of the increasingly complex language being read and (ii) how the complexification impacts the user's comprehension, and consequently the writing continuing the challenging texts. In principle, the system could then be extended to adapt the subsequent text challenges based on a combination of these form and meaning factors. An empirical evaluation the effectiveness of such a setup is reported in the next chapter.

Chapter 7

Complex Input Primed Writing—An Empirical ICALL Study Investigating Krashen's *i + 1*

Chapter highlights

What is already known about this topic:

- According to the Input Hypothesis (IH), input that is a little bit beyond the current level of the learner's interlanguage, or $i+1$, is optimal for second language acquisition.
- The implementation of the IH requires assessment of the learner's current proficiency level, the difficulty of the learning input, as well as a way to link the two spaces.
- Linguistic complexity analysis has been used to assess the readability of input texts and gauge learner proficiency. The two spaces are linkable with the multidimensional complexity vector distances (cf. Chapter 5).

What this study adds:

- The IH is operationalized with the complexity construct and made empirically testable.
- We investigate the optimal level of +1 challenge from the syntactic complexity perspective with a randomized control experiment implemented in an ICALL environment.

- The ICALL system used in the study demonstrates the effectiveness of individualized comprehensible input on proficiency development.

Implications for theory, policy, or practice:

- For theory: The IH is empirically confirmed with our intervention study.
- For practice: An ICALL system implementing the CIPW task based on the IH is effective in promoting L2 complexity development.

Abstract

Since its formulation by Krashen in the 1980s, the Input Hypothesis (IH), or *i + 1*, has gained general acknowledgment in the field of SLA. However, the theory has also drawn major criticism on its operationalizability and empirical testability. Despite the fact that the notion of *i + 1* has been in existence for more than three decades, the lack of empirical research on the hypothesis still limits its applicability in actual teaching practice. The current study tries to fill this gap by operationalizing the theory with linguistic complexity, which has been applied to analyzing learning input and learner productions. The purpose of the study is to investigate how input of different challenge levels (how much should the *+1* be) with regard to the linguistic complexity of the learners' production (the *i*) would result in the development of their L2 proficiency. A dedicated experimental system that implements the Complex Input Primed Writing scheme was developed to automatically select appropriate texts at four challenge levels (zero, low, medium, and high) based on the complexity of the learners' L2 production for continuation writing tasks. Results show that most students were able to make improvement matching the challenge they received given that the challenge was at the low and medium levels. The study essentially operationalized the IH and demonstrated an ICALL system that is capable of providing individualized and adaptive materials for L2 learning.

Related publication

This chapter is based on the following submitted manuscript:

- Chen, X., Meurers, D., and Rebuschat, P. (Submitted-a). Investigating Krashen's *i + 1*: An experimental ICALL study on the development of L2 complexity.

7.1 Introduction

In SLA, it is generally acknowledged that input at an appropriate difficulty level to the learner's proficiency level is optimal for L2 acquisition. This kind of input is *comprehensible*, as Krashen (1985) puts it in his IH. Comprehensible input needs to be neither too easy nor too difficult, but at a level that is a little bit beyond the current level of the learner's interlanguage, or $i + 1$. However, major criticism of the theory is that it is hard to operationalize the hypothesis and empirically test its validity (Ellis, 1990). As a result, the applicability of the theory into actual L2 teaching practice is greatly limited. In the current study, we tried to operationalize the IH with the complexity construct, which has been widely used in SLA studies to analyze both learning input and learner productions.

The constructs of linguistic complexity in general and syntactic complexity in particular have been widely used in SLA research to (a) gauge language proficiency, (b) assess production quality, and (c) benchmark language development (Ortega, 2012; Lu and Ai, 2015; Chen and Meurers, 2017a). They are either used as *independent* variables to predict text readability (Vajjala and Meurers, 2012; Collins-Thompson, 2014), evaluate writing quality (Taguchi et al., 2013; Ferris, 1994) and the like, or as *dependent* variables to investigate the effects of different learning tasks on language productions (Ellis and Yuan, 2004; Révész et al., 2017; Ong and Zhang, 2010) as well as to characterize writings by learners of different developmental stages (Vyatkina, 2013, 2012; Bulté and Housen, 2018) and/or with different backgrounds (Lu and Ai, 2015). Complexity has proved to be an effective construct for this research. However, most of the previous studies analyzed the linguistic complexity of either learning input or learner productions separately. Although there is already evidence showing that the input and production spaces of SLA is relatable with linguistic complexity (Chen and Meurers, 2018a), there has never been intervention studies exploring the effects of complex input on learner language development.

Although the goal of SLA is not to produce complex language as such, as their proficiency increases, second language (L2) learners usually demonstrate better mastery and more frequent use of complex language because of their expanding linguistic repertoire and capacity to use a wider range of the linguistic resources offered by the L2's grammar (Ortega, 2015). Foster and Skehan (1996) also characterized language development as 'progressively more elaborate language' and 'greater variety of syntactic patterning' (p. 303). Thus it is justifiable to use complexity to gauge the development of the learners' L2, or rather, as a proxy to their L2 proficiency. If we consider mastery of more complex language, which is manifested as the ability to understand and produce it appropriately, as L2 development, the question is then how

this development can be better promoted. Krashen's Input Hypothesis (1985) seems to provide a theoretical answer to the question. However, an empirical experiment will provide a more convincing and concrete answer.

Combining the theory of comprehensible input and complexity research in SLA, it can be hypothesized that the complexity of the target language input should then be slightly higher than what the learner can understand or produce—the current developmental stage of the learner. As was discussed previously, linguistic complexity can be used as a proxy to development/proficiency. It can also be used to assess the appropriateness of the input in terms of ease of understanding for learners of certain proficiency levels, i.e. to assess the readability of reading texts. However, it is still unclear whether the input chosen with the complexity analysis approach can practically promote the development of the learner's L2 and if it does, what should the optimal amount of complexity difference between the input and the learner production be. In other words, how big should the '+1' difference be. These are the empirical questions the current study tries to answer.

In what follows, we will first review previous research on the development of L2 complexity and how it can be related to the complexity of learning input to justify for the need of an intervention study to better promote complexity/proficiency development. Then the design of the experiment will be laid out and its results reported. The effects of different levels of input complexity, or *challenge* to the learner, on the development of their production complexity, or *improvement* will be investigated and discussed. We conclude that complex input as relates to L2 proficiency is effective in eliciting complex output to a certain extend: low and medium levels of challenge would result in improvement reaching the levels of the challenge, while higher challenge would result in failed attempt to catch up. In essence, the study operationalized the IH by unifying the learning input and learner production spaces with automatic analysis of their complexity. The results of the study form the basis of potential Intelligent Computer Assisted Language Learning (ICALL) systems to promote L2 proficiency with combined reading and writing tasks.

7.2 L2 complexity development

Ample empirical evidence from the SLA literature has shown that there is a strong correlation between the complexity of the learners' L2 production and the developmental stages they are at or their L2 proficiency (Bulté and Housen, 2018; Ortega, 2003; Wolfe-Quintero et al., 1998). Both lexical and syntactic complexity have been shown to correlate with L2 development in cross-sectional and longitudinal studies

(e.g. Verspoor et al., 2012; Vercellotti, 2017; Larsen-Freeman, 2006; Vyatkina, 2012; Vyatkina et al., 2015; Crossley and McNamara, 2014; Lu, 2011; Bulté and Housen, 2018). For instance, Verspoor et al. (2012) analyzed a corpus of 437 writings by Dutch learners of English as an L2 whose proficiency levels ranged from A1.1 to B1.2 according to the Common European Framework of Reference (CEFR) and found that most of the lexical and syntactic complexity measures they used, such as the Guiraud index, lexical sophistication, and mean length of T-unit and so on were able to distinguish between proficiency levels of writing expertise.

In particular, earlier studies have consistently shown that lexical sophistication measures, which are calculated by looking up the frequency of words used in a writing from some normed reference frequency lists, are revealing of the learners' proficiency levels (Jarvis et al., 2003; Laufer and Nation, 1995; Verspoor et al., 2008). Proficiency was found to be able to explain 46% of the variance in the initial scores of lexical variety in Vercellotti's (2017) study. In terms of syntactic complexity, Lu (2011) found that 10 out of the 14 measures he tested were able to discriminate proficiency levels. These measures include production length, complex nominal, and coordinate phrase indexes such as mean length of clause, complex nominals per T-unit, coordinate phrases per clause, etc. Crossley and McNamara (2014) also observed that 'a significant growth in syntactic complexity occurred in L2 writers as a function of time spent studying English' (p. 66).

Although the data used in the studies cited in the previous paragraph were mostly cross-sectional, longitudinal research also showed that lexical and syntactic complexity of L2 production would increase over time within individual learners. For example, Larsen-Freeman (2006) followed a group of five adult learners of English over a period of half a year by asking them to complete four narrative writing tasks with intervals of six weeks between consecutive writings. It was found that both lexical and syntactic complexity of the group increased steadily over time, although great inter- and intra-individual variability was observed on the individual level. Vyatkina (2012) also found a general upward trend on both the general and more specific syntactic complexity measures from her longitudinal L2 German data, corroborating Larsen-Freeman's (2006) findings. Another longitudinal study by Vyatkina and her colleagues (Vyatkina et al., 2015) further confirmed the effectiveness of syntactic complexity measures as developmental indices of proficiency levels.

This evidence supports the claim that linguistic complexity of L2 production increases with the development of the learner's target language proficiency. However, it is still difficult for researchers and language instructors to imply from these studies how to better promote the learners' proficiency from the complexity point of view.

Firstly, most of the previous studies were observational in nature. The learner production corpora were collected either cross-sectionally from learners of different grades or courses targeting learners of different proficiency levels (e.g. Ortega, 2003; Lu, 2011), or longitudinally at different time points (e.g. Vyatkina et al., 2015; Bulté and Housen, 2018). As a result, the change in complexity was always attributed to the time, grade, or course levels in which the L2 production was collected. Thus only a blanket explanation of instructional effect to complexity development is possible in most studies. It is difficult for researchers or language teaching practitioners to tease apart the actual cause of this development. However, as practitioners and L2 instructors, the more interesting question is usually how this research can benefit actual teaching practice. That is, how L2 teachers can help their students to better promote their L2 proficiency, or equivalently, to better use appropriately more complex language in the L2?

Secondly, previous studies have also consistently made evident that the development of complexity is characterized by variability and change (Lowie and Verspoor, 2015; Verspoor and Dijk, 2012). The linguistic complexity of L2 production does not always increase in parallel with proficiency, nor is it linear, constant, or guaranteed for all layers and sub-dimensions (Bulté and Housen, 2018). This developmental pattern makes it hard for L2 instructors to estimate the developmental stages a learner is in, so it would also be difficult for them to choose comprehensible learning materials to help the student move forward. As was discussed earlier, the choice of appropriate learning input should be based on accurate evaluation of the current proficiency of the learners. However, the 'non-linear waxing and waning' (Larsen-Freeman, 2006; Ortega, 2015) of their complexity developmental trajectories poses a threat to the accuracy of such evaluation at any single time point. A possible solution to this problem is to do online assessment of the learners' production dynamically so as to provide each learner with adaptive input. The current experiment provides a prototypical implementation of such a system.

In order to provide L2 teachers with more concrete guidance on how to better promote the learners' complexity development or proficiency, there arises a need to research on the relationship between the complexity of L2 input and that of the learners' production. According to the IH (Krashen, 1985), comprehensible input is an indispensable part of language acquisition. Adopting the theory from the perspective of linguistic complexity, it is thus reasonable to believe that by providing learners with input that is at the $i + 1$ complexity level in relation to that of their production the language teachers will be able to help them better acquire the L2. Preliminary research by Chen and Meurers (2018a) has shown promising results

of the effects of complex input on the complexity improvement of the learners' L2 production (see also Chapter 5). We used a continuation writing corpus collected by Wang and Wang (2015) who asked a group of Chinese learners of L2 English to read two stories (one in Chinese, and the other in English) with endings removed and to continue writing the stories in their L2. We then calculated two indexes from the complexity of the input stories and the students' writings:

$$challenge = complexity(English\ input\ text) - complexity(baseline\ writing)$$

and

$$improvement = complexity(continuation\ writing) - complexity(baseline\ writing)$$

A baseline writing is a continued English writing after reading the story in Chinese, while a continuation writing refers to a continued English writing after reading the story in English, the L2 of the participants. Significantly high Pearson's correlation coefficients (ranging from .28 to .96) were observed between challenge and improvement for the majority of the complexity measures they used. Our study proves that it is viable to relate learning input to learner production with complexity measures. It provides an operationalizable implementation of Krashen's $i+1$ hypothesis. However, because of the limitation of the dataset, it is still unclear whether different levels of challenge would result in different learning effects and how to account for the gap between the receptive and active knowledge of the L2 (understanding and producing the language).

This leads to the purposes of the current study, which boil down to the following research questions:

1. Would it foster proficiency development if L2 learners are challenged with input that is more complex than what they are capable of producing? In other words, would the complexity of their L2 production increase consistently if learners are exposed to input that is higher in complexity than that of the original L2 production by the learners?

2. If yes, what are the effects of different challenge levels on proficiency development?

Answers to these questions will not only provide L2 teachers and practitioners with concrete guidance on how to select learning input based on the evaluation of the learners' current proficiency levels, but also provide insights into the design of ICALL

systems for L2 acquisition because all analysis proposed in this study are automatizable. We tried to answer the research questions with a fully automatic intervention study, whose design and procedure are introduced in the next section.

7.3 Methods

There is a general lack of intervention studies on the effects of complex input on proficiency development. Such a study would require assessment of the learners' current proficiency and assignment of input of various challenge levels. It also needs to single out a test measure (independent variable) and control for all possible confounding factors. The experimental task should not only ensure that the treatment is received by the learners but also be able to elicit production from them for the purpose of evaluating the effects of the treatment. Based on these considerations, a CIPW task was conceived for the purpose of the current study. It is based on the continuation writing task designed by Wang and Wang (2015) but adds some modules to automatically analyze the learners' L2 production and choose reading input based on the analysis.

7.3.1 Automatic analysis of linguistic complexity

The advantage of using linguistic complexity to assess L2 proficiency so as to locate input appropriate for promoting language development is that the whole process is automatizable, making it possible to provide learners with individualized and adaptive learning materials. A number of general purpose systems have been developed for extracting complexity measures from both learning input and learner productions (e.g. Lu, 2010, 2012; McNamara et al., 2014; Kyle and Crossley, 2015; Chen and Meurers, 2016b). The experiment system used in the current study is built on the basis of the Common Text Analysis Platform (CTAP, Chen and Meurers, 2016b, see also Chapter 2), which is capable of extracting large numbers of complexity measures from multiple lexical, syntactic, and discoursal levels. The complexity of the texts in the reading corpus was extracted and stored in a database beforehand, while the analysis of the participants' writings were done online to dynamically choose input texts for the next CIPW task.

7.3.2 The Complex Input Primed Writing tasks

A CIPW task is a task in which the participants are asked to complete an essay—a narrative story, an argumentative writing or any other genres deemed appropriate—

whose ending has been removed. We removed the last quarter of the sentences in each essay for this experiment, leaving the first 75% of the essay to be read and continued by the participants. The task takers are instructed to continue writing the essay in a way that completes the narration or argumentation as coherently as possible. It is a focus-on-meaning task from the participants' perspective but offers linguistic priming for the writing because the participants need to read and understand the remaining part of the essay in order to be able to continue writing it. In the current study, all essays to be continued are chosen based on two individualized criteria: (a) the baseline complexity of the previous writings by the participant, and (b) the treatment group the participant is in. In order to answer whether different levels of complexity challenge would result in different improvement, four treatment groups were used: zero-, low-, medium-, and high-challenge groups.

We adopted the syntactic complexity measure of Mean Length of T-unit (MLTU) as the treatment measure of the current study. A T-unit is a minimally *terminalbe* unit (hence the name T-unit) or the shortest grammatically allowable sentence which consists of 'one main clause with all subordinate clauses attached to it' (Hunt, 1965, p. 20). The MLTU measure has been consistently found to discriminate L2 proficiency levels and develop in a somewhat linear manner (Lu, 2011; Ortega, 2003; Bulté and Housen, 2018). Ortega (2003) suggested that a difference of 2 words in MLTU be statistically significant to differentiate two consecutive proficiency levels. As a result, the participants in the zero-, low-, medium-, and high-challenge treatment groups were assigned texts that were +0, +2, +4, and +6 words more than their baseline MLTU respectively. The baseline MLTU was calculated as the mean MLTU of all writings the participant had submitted to the experiment system. For example, if a participant had been able to produce writings with a mean MLTU of 10 words and was randomly assigned into the low-challenge treatment group, in the next CIPW task, she would receive texts with an MLTU of 12 words (the baseline of 10 words plus a low challenge of 2 words). If she had been assigned to the medium-challenge group, the texts she would receive in the next CIPW task would have an MLTU of 14 words instead.

In order to make sure that the texts controlled for MLTU assigned to the participants were comprehensible to them, the experiment system chose from the reading corpus only texts that met the aforementioned criteria, as well as were closest to the participants' earlier production in all other complexity dimensions—the *nearest neighbours* in the complexity vector space. The CIPW experiment system was able to extract 558 lexical, syntactic, and cohesion complexity measures from a text. Consequently, except for the treatment measure of MLTU, all the other 557

measures were first normalized and then used to calculate the Euclidean distance between the learner production and the texts in the reading corpus. The distance between two points p and q in an Euclidean n-space can be calculated with the Pythagorean formula:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

For each CIPW task, the 10 texts from the reading corpus that were closest to earlier learner productions (10 nearest neighbors) were offered as choices for the participants to continue writing. It is reasonably assumed that the input texts that are close to the learners' production in terms of linguistic complexity are comprehensible by them. As a result, depending on the complexity of their writings, each learner would receive highly personalized and adaptive CIPW tasks.

7.3.3 The reading corpus

The corpus of reading essays for the CIPW tasks was collected from Newsela[1], an American educational website featuring articles on various contemporary topics that target students of different grades and reading abilities. The Newsela website adapts each published story into five different reading levels (ranging from the second to the twelfth grades), including the original version of the story as the 'Max' level. The reason for using the Newsela corpus in our study is that it offers a broad spectrum of variability in the linguistic complexity of the texts. This is important for a system that offers individualized input based on the learners' proficiency.

The Newsela website offers essays in different genres, including news, narratives, argumentations and so on. However, for a controlled intervention experiment like the current study, it is important to control for genre because it has been found to affect the complexity of the learner's production in previous research (e.g. Beers and Nagy, 2009; Way et al., 2000; Yoon and Polio, 2017). For example, Yoon and Polio (2017) found that the complexity of learner-produced argumentatives is higher than that of narratives. As a result, the current study restricted the CIPW task genre to argumentative writings. Six-hundred-thirty-five texts were obtained from the 'Opinion' and 'Pro/Con' sections of the Newsela website. The MLTU of these texts ranged from 7.42 to 30.42 (M = 14.60, SD = 4.04). Table 7.1 summarizes the profile of the Newsela corpus used in the present study.

[1] `https://newsela.com`

Grade Level	# Texts	# Words/Text	Mean MLTU	SD MLTU
2	16	427.25	8.24	0.59
3	99	502.20	9.56	0.82
4	46	704.43	11.12	0.90
5	105	804.48	12.58	1.01
6	72	956.83	14.45	1.11
7	80	993.35	15.50	1.27
8	46	1086.04	17.56	1.41
9	58	1032.05	18.33	1.74
12	113	1092.60	20.00	2.91

Table 7.1: Profile of the reading corpus used in the current study

7.3.4 Procedure

The experiment was conducted in a fully-automatic online environment specifically created for the study. After signing up to take part in the experiment, the participants received an email with login details to the experiment system, which was used to collect information on the participants' background, individual difference metrics[2], proficiency test, pre- and post-test writings, as well as the intervention treatment of 10 CIPW writings. The proficiency test used in the study was a web-adapted version of the C-tests (Klein-Braley, 1985) used by the Language Learning Center at the University of Tübingen for language course placement purposes. C-tests have been found to be predictive of general L2 proficiency (Dörnyei and Katona, 1992; Eckes and Grotjahn, 2006; Harsch and Hartig, 2016).

The difference between the pre-/post-test writings and the CIPW writings is that the former are free-writing tasks with only topic prompts, while the latter is a continuation writing task which provides the participants with the first three quarters of a text with which they are required to continue writing after reading it. The pre-/post-test writing topics are 'shared economy' and 'work from home' respectively. The selection of the first CIPW task essays is based on the complexity of the pre-test writing. The subsequent ones are based on the mean complexity of the submitted writings of individual participants. Figure 7.1 shows a screenshot of the experiment system, whose left navigation menu lists all the questionnaires and tasks the participants are required to complete in sequence.

[2]It should be noted that although we collected extensive data on the participants' motivation to learn a foreign language, their working memory and declarative memory, for the purpose of the current study, these data have not been used.

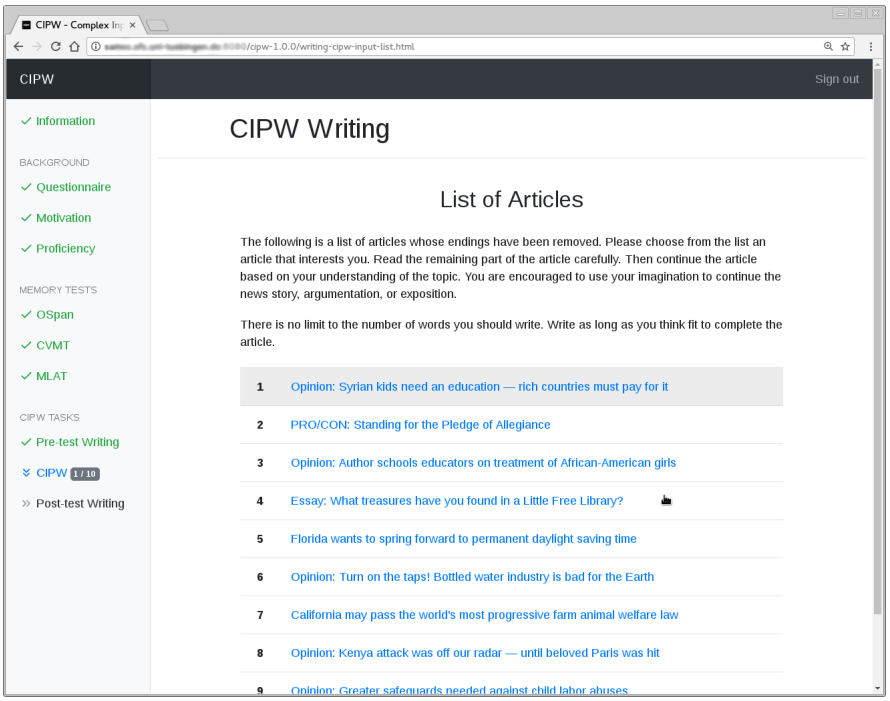

Figure 7.1: A screenshot of the CIPW experiment system

7.3.5 Participants

One-hundred-and-sixty-three Chinese learners of English from a Chinese university answered to the call-for-participation of the study and were sent login details to the experiment system. However, only 112 participants (68.7%) actually finished some writing tasks and were included in the analysis. The participants were randomly assigned into a control group and the four experimental groups who received different levels of challenge as described in the previous section. Participants in the control group did not do the CIPW writing tasks. Instead, they were required to finish the pre- and post-test writings with an interval of three weeks in between, which was also the time allowed for the other groups to finish all the writing tasks. On average, the number of writings each participant in the experimental groups finished was 8.68 ($SD = 4.35$).

Out of the 112 participants included in this analysis, 66 were male and 46 were female. Their ages ranged from 17 to 27 ($M = 18.98, SD = 0.88$). The mean number of years they had spent learning English was 10.56 years ($SD = 2.39$). In the background questionnaire, the participants were asked to self-indicate their English proficiency with a set of proficiency descriptors. Five participants thought they were *post-beginners*, 23 *lower-intermediates*, 67 *intermediates*, 15 *upper-intermediates*, and 2 did not report their proficiency. No participants considered themselves *beginners*. The C-test results also showed that the majority of participants were intermediate

learners of English. Table 7.2 summarizes the proficiency of the participants in each group.

Group	A1	A2	B1	B2	C1	C2	**Total**
Control	1	7	1	2	1	0	12
No-challenge	2	18	4	2	1	0	27
Low-challenge	2	12	7	2	1	1	25
Medium-challenge	0	15	4	4	1	0	24
High-challenge	1	10	5	6	2	0	24
Total	6	62	21	16	6	1	112

Table 7.2: Number of participants in each group and their proficiency distribution based on the C-test results

7.4 Results

In light of the research questions on whether complex input fosters L2 proficiency development and if it does, how much more complex the input should be with regard to the proficiency, we first present results on the comparison of the complexity of the pre- and post-test writings. Then we explored the patterns of writing complexity across time/tasks. The interaction between the complexity of input and proficiency development is operationalized as the longitudinal interaction between the challenge the participants received and the improvement they made from each CIPW task. Detailed analysis and results are reported in the following sub-sections.

7.4.1 Complexity of pre- and post-test writings

The variable of interest in the experiment is the MLTU of the participants' writings. On average, the participants were able to produce texts with a MLTU of 18.12 words ($SD = 6.5$) in the pre-test writing task. A one-way ANOVA confirmed that the five experiment groups did not differ in the MLTU of their initial writings ($F(4, 107) = .11, p \geq .1$). Out of the 112 participants who finished the pre-test writing, 71 of them finish the post-test writing. The mean MLTU of the post-test writings was 17.15 ($SD = 6.12$). No significant differences were found for the post-test writing MLTU among the groups either (One-way ANOVA $F(4, 66) = .91, p \geq .1$). The changes of MLTU between the pre- and post-test writings were calculated for those who finished the both writings. The mean changes were negligible ($M = .24, SD = 6.41$). Again, no significant differences were found among the experiment groups (One-way ANOVA $F(4, 66) = .43, p \geq .1$).

7.4.2 Developmental patterns of writing complexity

In order to observe how the complexity of the writings developed across CIPW tasks, an MLTU developmental trajectory was plotted for each individual participant. Figure 7.2 shows a typical developmental trajectory for each experimental group. The plots show a wavy developmental pattern for the complexity of the writings across tasks. For most participants, the complexity of their writings increased at the beginning of the experiment before falling to the beginning level and then back up again. Although there are individual differences in the magnitude (height and width) of the 'waves', the wavy pattern is observable in almost all participants who finished more than a few CIPW tasks.

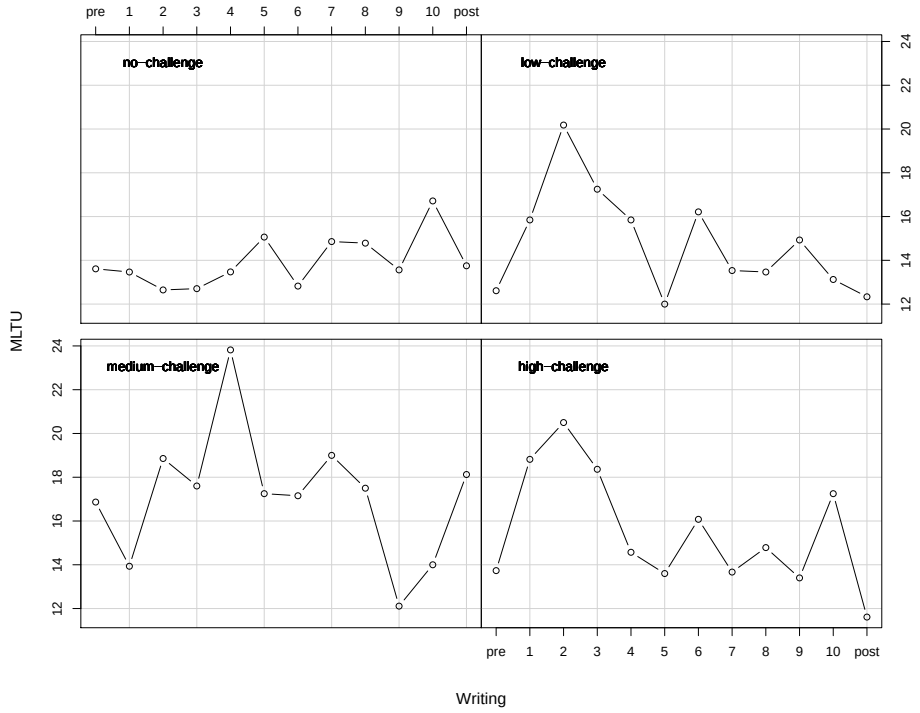

Figure 7.2: Example developmental trajectories of MLTU across writing tasks by participants of different experimental groups

7.4.3 Challenge and improvement

Following Chen and Meurers (2018a) and for the purpose of investigating the effect of input complexity on that of the continuation writings, the indexes of *challenge* and *improvement* were calculated. The original setup of the experiment system was to use the mean MLTU of all previous writings of the participant as the *baseline*. Depending on the groups the participants were assigned, they would always receive

texts that were challenging with respect to this baseline. As a result, the baseline would change dynamically as the experiment progressed because it would be re-calculated every time a new writing was submitted. If, for instance, a participant completed a writing with a higher MLTU than the mean MLTU of all her previous writings, the new mean MLTU would increase, resulting in a higher absolute MLTU for the next input as compared to that of the previous one, and vice versa. The baseline calculated in this way is called the *dynamic baseline* in our analysis. Another way to calculate the baseline is to use the mean MLTU of the pre- and post-test writings, which would result in a *static baseline* for each participant, because this baseline value does not change. Since the static baseline also depicts what the participant is capable of doing in terms of writing complexity without being primed by more complex input—it is calculated with the MLTU of the pre- and post-writings, it can also be considered as the *proficiency baseline* of the learner.

Equations 7.1 and 7.2 were used to calculate the challenge and improvement indexes, where C is the complexity measure, or MLTU in the case of the experiment. C_i denotes the complexity of input for a specific CIPW task, while C_w is the complexity of the participant's writing. The baseline complexity is denoted as C_{db} or C_{sb}, for dynamic and static baselines respectively.

$$challenge = C_i - C_{db/sb} \tag{7.1}$$

$$improvement = C_w - C_{db/sb} \tag{7.2}$$

Figures 7.3 and 7.4 plot the summarized relationship between the mean challenge the participants received and the average improvement they made. Figure 7.3 used the dynamic baseline, while Figure 7.4 used the proficiency baseline to calculate the plot indexes. Linear regression models were fitted for both calculations with challenge as predictor of improvement. As is also observable from the plots, challenge does not predict improvement when the indexes are calculated with the dynamic baseline (Figure 7.3, $R^2 = .03, F(1, 89) = 2.75, p > .1$). In contrast, the model with indexes calculated with proficiency baseline shows a clear linear trend: challenge is highly predictive of improvement (Figure 7.4, $\beta = .77, p \leq .01$; adjusted $R^2 = .45; F(1, 89) = 73.59, p \leq .01$). Comparison of the two models shows that the static proficiency baseline better helps explain the relationship between the complexity of the input texts and that of the continuation writings.

To further account for by-participant and by-task variation, a mixed-effect model was fitted with the `lme4` (Bates et al., 2015) package in R (R Core Team, 2015). The

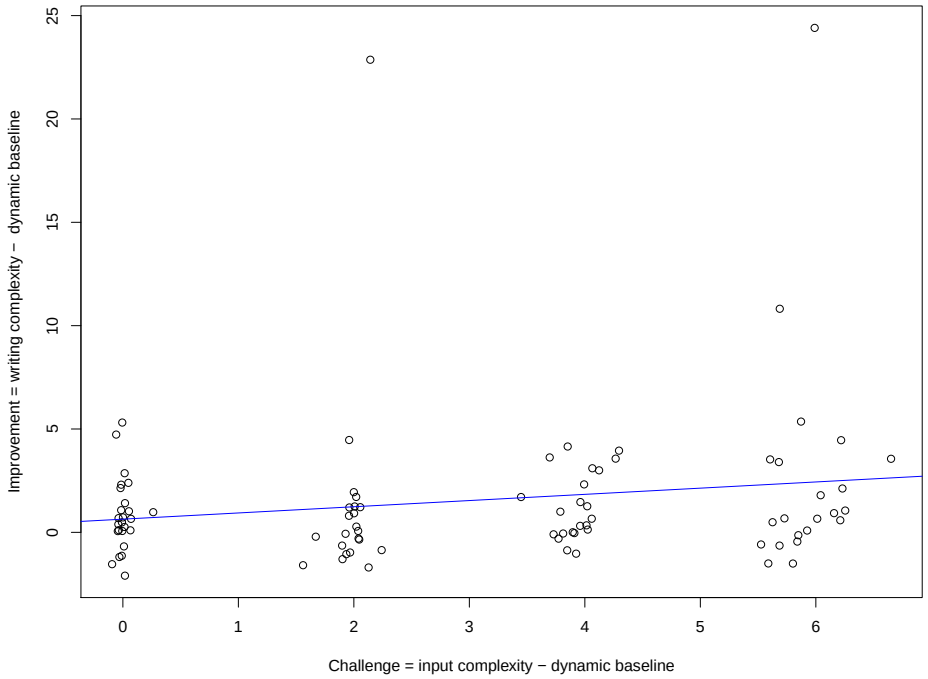

Figure 7.3: Mean improvement by challenge with dynamic baseline

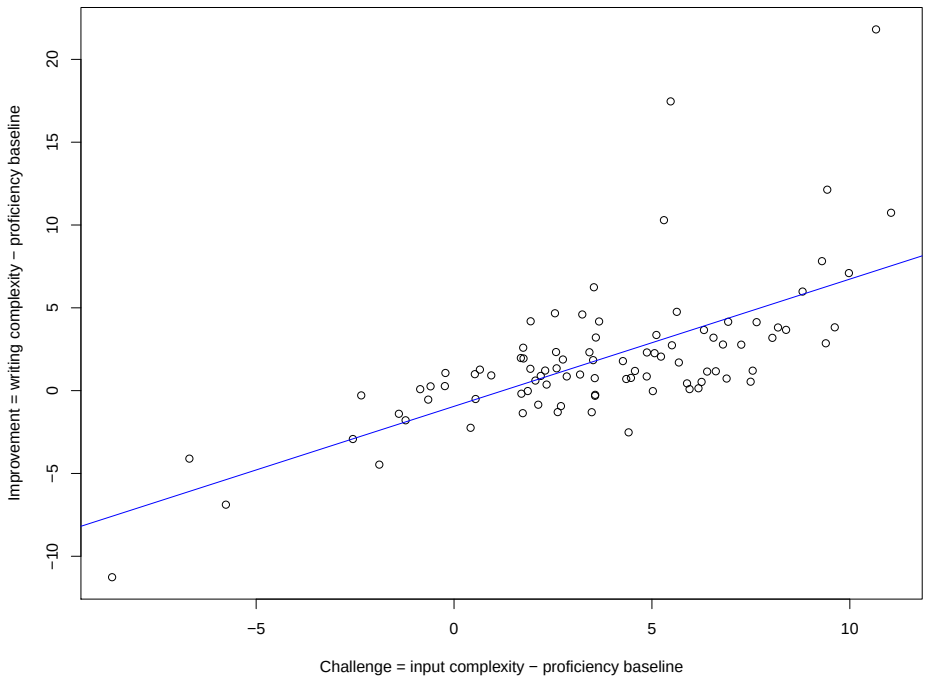

Figure 7.4: Mean improvement by challenge with proficiency/static baseline

challenge calculated with proficiency baseline was entered into the model as a fixed effect of complexity improvement in the participants' writings. As random effects, participants and the sequence of the writing tasks they completed were entered as both random intercepts and random slopes for the effect of improvement. Equation 7.3 shows the configuration of the mixed-effects model used in the R environment. Residual plots of the model showed no obvious violation of homoscedasticity or normality. No interaction was found between challenge and the proficiency of the participants as assessed by the C-test. Comparison of the models with and without such interaction was done with likelihood ratio tests and yielded $\chi^2 = 5.32, p \geq .1$. In the model denoted by Equation 7.3, 19% of the variance in improvement was explained by the fix terms (marginal $R^2 = .19$), while both the fixed and random factors were able to account for 72% of the same variance (conditional $R^2 = .72$).

$$imprv \sim chllng + (1 + chllng|sbjct) + (1 + chllng|wrtng) \tag{7.3}$$

Figure 7.5 shows some patterns of interaction between challenge with regard to proficiency baseline and the complexity improvement of the participants' writings after receiving the challenge. Participants from the same experimental groups were plotted on the same rows. Hence from bottom up, the rows of plot panels represent data from the no-, low-, medium- and high-challenge groups respectively. The columns of panels in Figure 7.5 show different interactional patterns. The left-most column, except the top panel, shows participants who were able to 'outperform' the challenge, hence the black solid lines are mostly above the blue dashed ones. The second column shows participants who were able to 'catch up' with the challenge, while the last two columns show participants who could barely meet the challenge or fell completely behind it. Each type of interactional pattern between challenge and improvement is observable in all experimental groups. However, the general trend is that the groups that received higher levels of challenge usually witness more cases of failures to achieve the same levels of improvement as the challenge.

7.4.4 Summary of main results

No significant difference was found between the complexity of the pre- and post-test writings, which were two free writing tasks on different topics. However, wavy developmental patterns were observable from most participants. The complexity of the CIPW writings fluctuated in response to the complexity of the input, but within a certain limit. The response of varying writing complexity to the complexity of the input could be captured with a linear mixed effects model, which found that the

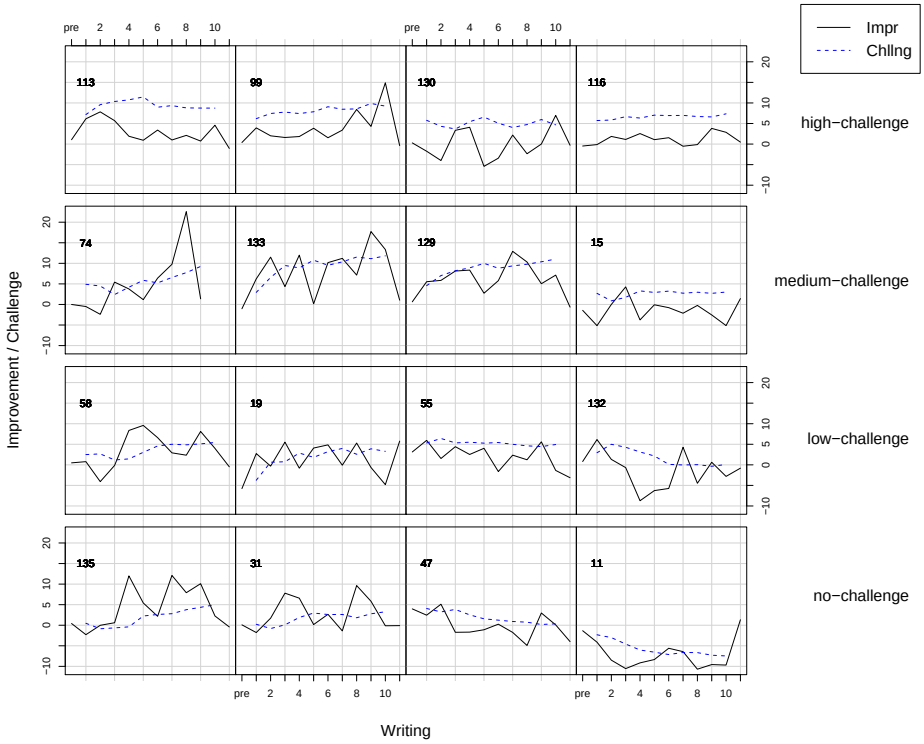

Figure 7.5: Sample patterns of interaction between challenge and improvement

complexity of the input in relation to the proficiency baseline (i.e. the challenge) was able to explain 19% of the variance in the improvement the participants made in the CIPW tasks. It was also found that in general, the participants were able to make improvement that matched with the low- or medium-level challenge. But more students would fail to catch up if the challenge is more than 2 levels higher than their proficiency baseline.

7.5 Discussion

The first purpose of the current study was to investigate the effects of complex input on the development of L2 proficiency, which was operationalized as the complexity of the learner's L2 production. A randomized control experiment methodology was adopted with the syntactic complexity measure of MLTU singled out as the treatment variable, while the other complexity measures were used to control for the general complexity of the input. Results from the CIPW experiment shows that regardless of the challenge the participants received during the 10 CIPW writing tasks, the complexity of their post-test writings did not differ significantly from the pre-test writings. This is a somewhat disappointing but also understandable result. In the current study, L2 proficiency was operationalized as the complexity of the

L2 learners' writing, and to be more specific, as the syntactic complexity measure of MLTU. Previous research has consistently found MLTU to be the most distinguishing complexity measure of L2 proficiency (Bulté and Housen, 2018). It was also found to develop somewhat linearly across proficiency levels (Lu, 2011; Ortega, 2003). The fact that the treatment of 10 CIPW tasks did not promote the development of MLTU in free writing tasks may be explained as no learning effect of the intervention. However, what we did observe from the developmental trajectories of the writing complexity across CIPW tasks (cf. Figures 7.2 and 7.5) was that the complexity of the participants' CIPW writings fluctuated in response to the complexity of the input. It would be expected that the increased production complexity be carried over to free writing tasks given these observations. The fact that it did not happen may also be due the length of the treatment. It could well be that 10 CIPW tasks are not enough to foster significant changes in proficiency, hence no significant changes in the pre- and post-test writings. Another possible reason for this finding is task-effect, which has been found to have large effects on the complexity of L2 productions (e.g. Robinson, 2011; Michel et al., 2007; Tabari, 2016; Yoon and Polio, 2017; Alexopoulou et al., 2017; Yuan and Ellis, 2003). Even if the learners' ability to use more complex language have been expanded, when they are not primed to do so as in the situation of the free post-test writing, they would still produce language at the 'comfort zone' of their proficiency, rather than maximizing their complexity potentials.

The wavy developmental pattern of writing complexity for most participants reported in Section 7.4.2 is not difficult to explain. Since the experiment was set up in a way that the complexity of the next CIPW input would be based on the mean complexity of all previous writings of a participant, the complexity of the input would increase if the participant submitted a more complex writing. The complexity of the input would keep growing if the participant was able to keep up with the challenge, until it reached a point when the participant failed to cope with the increased complexity level. The result would be lower complexity of the continuation writing as compared to that of the input, drawing the average complexity of all submitted writings down if the new submission has lower complexity than the mean of all previous submissions. If the participant still failed to keep up with the new complexity level, the system would lower the input complexity again automatically until it reached a level that the participant could catch up. Then the process repeats itself, hence the multiple waves across CIPW tasks. The result shows how the CIPW system is capable of dynamically adapting its tasks to the development of individual learners, which is desirable for ICALL systems.

As for the results on the relationship between challenge and improvement, it was found that only when the both indexes were calculated with the participants' proficiency baseline was challenge able to predict improvement. This suggests that the static proficiency baseline is a more accurate representation of the learners' L2 proficiency from the complexity perspective than the dynamic baseline. Combining with the finding that no difference was found between the pre- and post-test writing complexity, it can be concluded that the complexity improvement the participants gained during the course of the CIPW tasks should not be considered as promotion of proficiency levels. At least the ability to use more advanced language had not yet been integrated as part of the learners' stable proficiency. This stabilization may require more practice. Another explanation may be that the participants had already been able to produce the more complex language but due to task or other factors, they did not use the advanced language in the free writings tasks. They would do so only when they were primed by the more challenging input in the CIPW tasks.

With the proficiency baseline, even after controlling for participants and writing tasks, it was found that challenge was still able to explain 19% of the variance in improvement. These results suggest that although the participants were not explicitly informed about the characteristics of the input they received, they still adapted the complexity of their writings to that of the input. This phenomenon may be seen as a type of implicit learning and priming effect. Implicit learning happens when L2 learners do not pay conscious attention to meaning negotiation or sentence construction (Ellis, 2005). Structural, or syntactic priming refers to the tendency that a speaker is more likely to use the same syntactic structures over the alternatives as the ones they have been exposed to in recent discourse (Bock, 1986). Studies on L2 syntactic priming have found that L2 learners are more likely to advance to a higher stage in the developmental sequence if they are primed with developmentally more advanced forms (e.g. McDonough and Mackey, 2006, 2008; Shin and Christianson, 2012) because priming could strengthen knowledge representations (Nobuyoshi and Ellis, 1993) and make the retrieval of linguistic forms more proceduralized (de Bot, 1996). The challenging input that had a higher complexity level in the CIPW tasks can be considered as more advanced language with regard to the learners' proficiency. After being exposed to and primed by the complex input, participants receiving moderate levels of challenge were able to made improvement matching the challenge. As a result, we tend to believe that learning did occur after completing the CIPW tasks, although a proficiency promotional effect is yet to be observed in free writing tasks.

The analysis of the interaction between the challenge and improvement trajectories (Figure 7.2) suggests that learners vary in how they react to the complexity challenge. Some learners are capable of coping with both medium and low levels of challenge, while others struggle even with low challenge. However, most learners in the high-challenge group failed to catch up with challenge. This finding confirms the Input Hypothesis, or *i + 1* by Krashen (1985). But the CIPW experiment is the first to make the hypothesis concrete and empirically testable. The implementation of the automatic CIPW procedure makes it possible to transfer the findings in this study into practical ICALL systems.

7.6 Summary

The current study is built upon the previous finding that the spaces of L2 learning input and learner production is relatable by a common analysis of their complexity (Chen and Meurers, 2018a). While most previous studies on linguistic complexity tend to characterize input or production separately, our CIPW experiment was designed to bring the two aspects of L2 learning together. The main interests of the study were on whether the complexity of learning input would affect the production complexity, which was seen as a proxy to the proficiency of the learners' L2. Results from the experiment suggest that L2 learners can be implicitly primed by syntactically more complex, or more advanced language with respect to their proficiency as measured by the complexity of their L2 production. It is believed that this priming effect would lead to L2 learning and ultimately increased L2 proficiency, although this effect was yet to be detected in free writing tasks.

One contribution the study offers to SLA research is that it operationalized and empirically tested the widely-acknowledged Input Hypothesis, or *i + 1*, whose greatest criticism is its testability and operationalizability (Ellis, 1990). The study also added a new perspective for investigating linguistic priming—from the complexity point of view rather than the traditional lexical and syntactic perspectives. Last but not least, since the accurate analysis of linguistic complexity of both learning input and learner production is highly automatizable due to the latest development of computational linguistics and natural language processing, the automatic CIPW task procedure can be integrated into practical ICALL systems to provide L2 learners with individualized and adaptive learning opportunities. The traceable developmental trajectories of the learners' proficiency from such systems will also shed further light on the effectiveness and working mechanisms of implicit learning from the complex input.

It should be acknowledged that the study also suffers from a few limitations. Firstly, a single syntactic complexity measure was used as a proxy to L2 proficiency. Although MLTU has been found to be most predictive of L2 proficiency levels, it is over-simplistic to consider it as the whole of the proficiency construct. As hundreds of complexity measures have been devised by previous research on complexity, future research should also focus on the co-varying factors instead of individual measures. Another interesting direction is to also explore how the complexity factors interact with each other when they are used as complex input for L2 learning purposes, ideally also taking into account individual learner differences. Furthermore, we were unable to investigate the long term effect of complex input on the development of general L2 proficiency due to the limited number of CIPW treatments. Future research could also tackle this problem with more treatments and over a longer period of time.

Chapter 8

Conclusions

This dissertation presents studies revolving around the construct of complexity, which is a multidimensional construct that has been widely used in SLA research, but yet to be further addressed. Complexity is not only an important instrument for analyzing learning input and learner production, but has also developed into a major subject matter of SLA research, which has its own working mechanism, cognitive and psycholinguistic processes, and developmental dynamics. The studies in this dissertation tried to address the conceptualization, measurement, and application of complexity in L2 learning research by approaching the construct from the structural, cognitive, and developmental dimensions with theories and tools from the fields of NLP, ML and SLA.

The research started by focusing on the structural dimension of complexity. A review of previous research utilizing the complexity instrument revealed the need to address the multidimensionality of complexity with robust NLP technologies, especially when it comes to the analysis of large amount of natural language data. As a result, a Web-based system—the Common Text Analysis Platform—for supporting comprehensive measurement of linguistic complexity from multiple levels of linguistic representations was developed and released with free access for the research community. The CTAP system provides researchers with a convenient and efficient method to extract a comprehensive set of complexity measures from multiple linguistic levels. The modularized framework of the system makes it easy to create new feature extractors that are pluggable into the system, thus allowing for collaborative development and expansion. The system was also used in a few other studies reported in this dissertation, proving the usefulness and effectiveness of the system.

The research continued by zooming into lexical complexity for readability assessment, which was aimed at finding comprehensible input for language learning

purposes. The study investigated how the word-level feature of lexical frequency can be used to characterize the text-level readability. Three experiments tested the effectiveness of different frequency norms, different types of frequency measures, and different approaches to use the frequency values in predicting text readability. Results show that lexical frequency is highly predictive of text readability. The best predictive models are constructed with frequency lists that best represent the exposure users of the language experience, because such lists are more likely to reflect the vocabulary retrieval and perception cognitive load of the readers. Depending on the requirements of practical applications, one could choose either the simple but efficient model with frequency mean and SD, or a more sophisticated but also more accurate model with the clustering frequencies as features of the readability model. It is concluded that although a single measure of lexical complexity is capable of achieving a relatively high level of estimation accuracy, a proper characterization of text readability should still be done with a more comprehensive set of complexity measures which take into account the morphological, lexical, syntactic, cohesive and cognitive demand aspects of complexity.

Complexity analysis was further applied to address a controversial issue of whether there is trade-off between complexity and the other two common descriptors of learner language—accuracy and fluency. Conflicting theories have been proposed and both have found empirical support in previous studies. Our study adds to the discussion by adopting a comprehensive view of complexity instead of the more limited reductionistic view adopted by most of the other studies. The reductionistic view of complexity is argued to be the source of the disputes on the issue. Our study also used longitudinal data from natural L1 and L2 instructional settings instead of elicited data from controlled experimental settings, making the conclusions more convincing. For both the L1 and L2 groups, complexity and accuracy both showed development over a four-month period. However, while the L2 group's development was mainly at the lexical and morpho-syntactic levels, the development of L1 group was mainly at the discourse levels. No evidence was found for a competing relation, or trade-off, between complexity and accuracy. Practically, these results suggest that it is not necessary for language teachers to prioritize complexity or accuracy in their classes—they can develop simultaneously. Furthermore, for courses targeting learners of different proficiency levels, the focus of instruction should be adjusted to the learner's abilities to support the development of different linguistic areas.

It is generally acknowledged that learners need authentic target language input that matches their language ability, or comprehensible input to promote language acquisition. The selection of this kind of input requires assessment of both the

input and the learner's ability to understand the language. Complexity analysis of learning input and learner production has proved to be successful in both respects. However, a successful ICALL system would require the unification of the input and ability spaces so as to be able to provide the learner with developmental input selection automatically. Complexity feature vector distance was proposed as a link of the readability and proficiency spaces and its validity was tested empirically. It was concluded that complexity feature vector distance can not only relate the two spaces at an aggregate level, but also at the fine-grained individual linguistic complexity feature level.

Based on these findings, the ICALL system SyB was developed. By calculating the distances between the student production and the texts in the reading corpus, the system is capable of selecting reading texts that match the student's overall language ability or a specific aspect of their ability. Students are free to adjust the overall or specific complexity levels of the input selected for them. The system showcases how the fields of SLA and NLP can be combined to develop useful applications to assisted language learning based on solid theoretical and empirical research findings.

In order to understand the effects of comprehensible input selected with the complexity feature vector distance approach on L2 development, based on previous research on continuation writing, we developed the CIPW task. The purpose of the task was to find out whether different levels of input challenge would result in different levels of L2 development, and if they do, what would the optimal level of challenge for second language acquisition purposes be. An intervention study was conducted to answer these questions. It was found that most students were able to make improvement at a level matching the level of challenge they received if the challenge was at the low or medium levels. Essentially, this study operationalized and validated the $i+1$ hypothesis, a classic but hard to implement theory of SLA.

In all, this research has focused on the conceptualization, measurement, and application of complexity. It tackled all the major dimensions of the construct, including the structural, cognitive, and developmental dimensions. Although we have gained a lot of new insights into complexity, its interrelationship with the other language descriptors, and the potential applicability of the construct in language acquisition research and practice, there are still unanswered questions and new areas to explore. For example, in terms of measurement, more needs to be done to provide automatic complexity analysis systems for languages other than English. Although systems for extracting complexity features from other languages exist, making these systems easily accessible by the research community is yet to be realized. With regard to the application of complexity in SLA research, future work

should focus more on how to improve the quality of language assessment with linguistic complexity for various purposes. For example, for assessing text readability for comprehensible input selection, more comprehensive models utilizing the full set of complexity features should be constructed based on findings about how to better use the individual features. Better prediction models could be achieved only by taking the multidimensionality of the complexity construct into consideration because every dimension could potentially contribute to comprehension difficulty for learners with different characteristics. Future research could also focus on the long term effects of complex input on language development. Although our CIPW experiment showed improvement in writing complexity matching the complex input, no transfer effect was detected in free writing tasks, probably due to the lack of enough treatment cycles. It is thus difficult to conclude on the effects of complex input on proficiency development. This question is only answerable with longer intervention studies, preferably also taking into account individual learner differences, such as working memory, motivation, learning strategies, and/or cognitive and learning styles.

Limitations of the current research notwithstanding, this dissertation has contributed significantly to the understanding, measurement, and application of the complexity construct in SLA research. It provides a new ground for approaching L2 learning and instruction by combining state-of-the-art NLP and other computer technologies with empirically validated language acquisition theories. We strongly believe that this new ground will lead to a promising avenue for future research in SLA, ICALL, and the other applied linguistics fields of study in general.

Bibliography

Aarts, J. and Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In Granger, S., editor, *Learner English on Computer*, pages 132–141. Longman, New York.

Adelman, J. S., Brown, G. D. A., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science*, 17:814–823.

Ahmadian, M. J. and Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, 15(1):35–39.

Ai, H. (2017). Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. *ReCALL*, 29(3):313–334.

Ai, H. and Lu, X. (2013). A corpus-based comparison of syntactic complexity in nns and ns university students' writing. In Díaz-Negrillo, A., Ballier, N., and Thompson, P., editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 249–264. John Benjamins, Amsterdam.

Alexopoulou, T., Michel, M., Murakami, A., and Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1):180–208.

Amaral, L. and Meurers, D. (2008). From recording linguistic competence to supporting inferences about language acquisition in context: Extending the conceptualization of student models for intelligent computer-assisted language learning. *Computer-Assisted Language Learning*, 21(4):323–338.

Amaral, L., Meurers, D., and Ziai, R. (2011). Analyzing learner language: Towards a flexible NLP architecture for intelligent language tutors. *Computer-Assisted Language Learning*, 24(1):1–16.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Balota, D. A. and Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, 10:640–357.

Bardovi-Harlig, K. and Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11(1):17–34.

Barkaoui, K. (2014). Quantitative approaches for analyzing longitudinal data in second language research. *Annual Review of Applied Linguistics*, 34:65–101.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beers, S. F. and Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality:which measures? which genre? *Reading and Writing*, 22(2):185–200.

Beneš, E. (1976). Syntaktische besonderheiten der deutschenwissenschaftlichen fachsprache. In *Fachsprachen. Terminologie, Struktur, Normung*, volume 4, pages 88–98. DIN Deutsches Institut für Normung e.V.

Benjamin, R. G. (2012). Reconstructing readability: recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Bentler, P. M. and Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, 34(2):181–197.

Bernhardt, E. B. and Kamil, M. L. (1995). Interpreting relationships between l1 and l2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, 16:15–34.

Bitchener, J. and Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, 19(4):207–217.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33(1):1–17.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Bricker, P. D. and Chapanis, A. (1953). Do incorrectly perceived stimuli convey some information? *Psychological Review*, 60:181–188.

Brown, C. G. (2002). Inferring and maintaining the learner model. *Computer Assisted Language Learning*, 15(4):343–355.

Brown, R. (1973). *A First Language*. Harvard University Press, Cambridge.

Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41(4):977–990.

Bull, S. (1994). Learning languages: Implications for student modelling in ICALL. *ReCALL*, 6(1):34–39.

Bulté, B. and Housen, A. (2012). Defining and operationalising L2 complexity. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 Performance and Proficiency*, pages 21–46. John Benjamins.

Bulté, B. and Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26:42–65.

Bulté, B. and Housen, A. (2018). Syntactic complexity in l2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, 28(1):147–164.

Bulté, B., Housen, A., Pierrard, M., and Van Daele, S. (2008). Investigating lexical proficiency development over time—the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies*, 18(3):277–298.

Bygate, M. (1999). Quality of language and purposeof task: patterns of learners' language on two oral communication tasks. *Language Teaching Research*, 3(3):185–214.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS: Basic Concept, Applications, and Programming*. Lawrence Erlbaum Associates, Mahwah, NJ.

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education*, 20(1):50 – 66.

Bytgate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In Willis, J. and Willis, D., editors, *Challenge and change in language teaching*, pages 136–146. Heinemann, London.

Cahill, A. (2015). Parsing learner text: to shoehorn or not to shoehorn. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 144–147, Denver, Colorado, USA. Association for Computational Linguistics.

Carroll, J. B., Davies, P., and Richman, B. (1971). *Word Frequency Book*. Houghton Mifflin, Boston.

Casanave, C. P. (1994). Language development in students' journals. *Journal of Second Language Writing*, 3:179–201.

CCSSO (2010). Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects. Technical report, National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

Chall, J., Conard, S. S., and Harris-Sharples, S. (1991). *Should Textbooks Challenge Students? The Case for Easier and Harder Books*. Teachers College Press, New York.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12(3):267–296.

Chapelle, C. A. and Heift, T. (2009). Individual learner differences in CALL: The FID construct. *CALICO Journal*, 26(2):246–266.

Chen, E. and Cheng, E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, 12(2):94–112.

Chen, L., Tokuda, N., and Xiao, D. (2002). A POST parser-based learner model for template-based ICALL for Japanese-English writing skills. *Computer Assisted Language Learning*, 15(4):357–372.

Chen, X. and Meurers, D. (2016a). Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications at NAACL*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

Chen, X. and Meurers, D. (2016b). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity at COLING*, pages 113–119, Osaka, Japan, 11th December. The International Committee on Computational Linguistics.

Chen, X. and Meurers, D. (2017a). Challenging learners in their individual Zone of Proximal Development using pedagogic developmental benchmarks of syntactic complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa*, pages 8–17, Gothenburg, Sweden, 22nd May. Linköping University Electronic Press, Linköpingsuniversitet.

Chen, X. and Meurers, D. (2017b). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, Advance Online Access.

Chen, X. and Meurers, D. (2018a). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, In press.

Chen, X. and Meurers, D. (2018b). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Chen, X., Meurers, D., and Rebuschat, P. (Submitted-a). Investigating Krashen's *i + 1*: An experimental ICALL study on the development of L2 complexity.

Chen, X., Weiß, Z., and Meurers, D. (Submitted-b). Is there a developmental trade-off between complexity and accuracy in L1 and L2 acquisition?

Chinkina, M. and Meurers, D. (2016). Linguistically-aware information retrieval: Providing inputenrichment for second language learners. In *Proceedings of the*

11th Workshop on Innovative Use of NLP forBuilding Educational Applications, pages 188–198, San Diego, CA.

Chinkina, M. and Meurers, D. (2017). Question generation for language learning: From ensuring texts are read to supporting learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 334–344, Copenhagen, Denmark.

Choi, I.-C. (2016). Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning*, 29(2):334–364.

Chomsky, N. (1986). *Knowledge of language*. Praeger, New York, NY.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, 11:38–63.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, 26(2):227–248.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of past, present, and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11(4):367–383.

Crossley, S., Dufty, D., McCarthy, P., and McNamara, D. (2007). Toward a new readability: A mixed model approach. pages 197–202.

Crossley, S., Greenfield, J., and McNamara, D. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

Crossley, S. A., Kyle, K., andLiang Guo, L. K. A., and McNamara, D. S. (2014). Linguistic microfeatures to predict l2 writing proficiency: Acase study in automated writing evaluation. *Journal of Writing Assessment*, 7(1):1–15.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2015). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016). The development and use of cohesive devices in l2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32(Supplement C):1–16.

Crossley, S. A. and McNamara, D. S. (2012). Predicting second language writing proficiency: The roles ofcohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Crossley, S. A. and McNamara, D. S. (2014). Does writing development equal writing quality? a computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26:66–79.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies*, 1(2):1–23.

Dale, E. and Chall, J. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27(Jan. 21 and Feb. 17):1–20, 37–54.

Dale, E. and Chall, J. (1949). The concept of readability. *Elementary English*, 26(3).

de Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language Learning*, 46(3):529–555.

De Clercq, B. and Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in l2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, 101(2):315–334.

De Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, 19(2):249–276.

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., Singh, A., Shankar, Y., Pulickal, L., Rajkumar, A., and Parmar, N. J. (2016). Tacit: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, pages 1–10.

DeKeyser, R. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In Doughty, C. and Williams, J., editors, *Focus on Form in Classroom Second Language Acquisition*, pages 42–63. Cambridge University Pres.

Derwing, T. M. and Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of l2 accented speech. *Applied Language Learning*, 13:1–18.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25(2):108–127.

Dion, P. A. (2008). Interpreting structural equation modeling results: A reply to martin and cullen. *Journal of Business Ethics*, 83(3):365–368.

Dörnyei, Z. and Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing*, 9(2):187–206.

DuBay, W. H. (2006). *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

Dugast, D. (1979). *Vocabulaire et stylistique*. I Théâtre et dialogue [Vocabulary and style. Vol. 1 Theatre and dialogue]. Slatkine-Champion, Geneva, Switzerland.

Eckes, T. and Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3):290–325.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, 27(2):305–352.

Ellis, R. (1990). *Instructed Second Language Acquisition*. Blackwell, Oxford.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press, Oxford.

Ellis, R. and Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford University Press, Oxford.

Ellis, R. and Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26(1):59–84.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of esl compositions. *Journal of Second Language Writing*, 4:139–155.

Evans, N. W., Hartshorn, K. J., Cox, T. L., and de Jel, T. M. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing*, 24:33–50.

Evans, N. W., Hartshorn, K. J., McCollum, R. M., and Wolfersberger, M. (2010). Contextualizing corrective feedback in second language writing pedagogy. *Language Teaching Research*, 14(4):445–463.

Feng, L. (2010). *Automatic Readability Assessment*. Doctoral dissertation, City University of New York (CUNY).

Feng, L., Elhadad, N., and Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 229–237, Athens, Greece. Association for Computational Linguistics.

Ferraris, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 Performance and Proficiency*, pages 277–298. John Benjamins.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, 28(2):414–420.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Flor, M. and Klebanov, B. B. (2014). Associative lexical cohesion as a factor in text complexity. *International Journal of Applied Linguistics*, 165(2).

Flor, M., Klebanov, B. B., and Sheehan, K. M. (2013). Lexical tightness and text complexity. In *Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pages 29–38, Atlanta, Georgia. Association for Computational Linguistics.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307.

Foster, P. and Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(3):299–323.

François, T. and Fairon, C. (2012). An "AI readability" formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 466–477, Stroudsburg, PA, USA. Association for Computational Linguistics.

François, T. and Watrin, P. (2011). On the contribution of mwe-based features to a readability formula for french as a foreign language. In *Proceedings of Recent Advances in Natural Language Processing*, pages 441–447, Hissar, Bulgaria.

Freed, B. (1995). What makes us think that students who study abroad become fluent? In Freed, B., editor, *Second language acquisition in a study abroad context*, pages 123–148. John Benjamins, Amsterdam.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: TheEF-Cambridge open language database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*. Cascadilla Press.

Ghadirian, S. (2002). Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning & Technology*, 6(1):147–164.

Ghasemi, M. (2013). An investigation into the use of cohesive devices in second language writings. *Theory and Practice in Language Studies*, 3(9):1615–1623.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.

Göpferich, S. and Neumann, I. (2016). Writing competence profiles as an assessment grid? – students' L1 and L2 writing competences and their development after one semester of instruction. In *Developing and Assessing Academic and Professional Writing Skills*, pages 103–140. Peter Lang, Bern, Switzerland.

Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., and Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, 19(3).

Granger, S. and Rayson, P. (1998). Automatic profiling of learner texts. In Granger, S., editor, *Learner English on Computer*, pages 119–131. Longman, New York.

Grant, L. and Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2):123–145.

Gray, W. S. and Leary, B. E. (1935). *What makes a book readable*. University of Chicago Press, Chicago.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie [The statistical characteristics of vocabulary: An essayin methodology]*. Presses universitaires de France, Paris.

Hancke, J., Meurers, D., and Vajjala, S. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Harsch, C. and Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, 33(4):555–575.

Haseley, L. (1957). *The relationship between cue-value of words and their frequency of prior occurrence*. Unpublished master's thesis, Ohio university.

Hawkins, R. (2001). *Second Language Syntax*. Blackwell, Oxford.

Heift, T. (2004). Corrective feedback and learner uptake in call. *ReCALL*, 16(2):416–431.

Heilman, M. J., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT 2007*, pages 460–467, Rochester, NY. Association for Computational Linguistics.

Hennig, M. and Niemann, R. (2013). Unpersönliches schreiben in der wissenschaft: Einebestandsaufnahme. *Informationen Deutsch als Fremdsprache*, 4:439–455.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *The Modern Language Journal*, 80(3):309–326.

Housen, A. (2014). Difficulty and complexity of language features and second language instruction. In Chapelle, C., editor, *The Encyclopedia of Applied Linguistics*, pages 2205–2213. Wiley-Blackwell.

Housen, A. (2015). L2 complexity—a difficult(y) matter. Oral presentation given at the Measuring Linguistic Complexity: A Multidisciplinary Perspective workshop, Université catholique de Louvain, Louvain-la-Neuve.

Housen, A., Daele, S. V., and Pierrard, M. (2005). Rule complexity and the effectiveness of explicit grammar instruction. In Housen, A. and Pierrard, M., editors, *Investigations in instructed second language acquistion*, pages 235–270. Mouton de Gruyter, Berlin.

Housen, A., Kuiken, F., and Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 1–20. John Benjamins.

Housen, A., Kuiken, F., Zuengler, J., and Hyland, K. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Howes, D. H. and Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41:501–410.

Hulstijn, J. and de Graff, R. (1994). Under what conditions does explicit knowledge of a second languagefacilitate the acquisition of implicit knowledge? a research proposal. *AILA Review*, 11:97–112.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Hunt, K. W. (1965). Grammatical structures written at three grade levels (research report no. 3). Technical report, National Council of Teachers of English, Champaign, IL.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of swedish. *Journal of Multilingual and Multicultural Development*, 9:67–84.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4(1):51 – 69.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63:87–106.

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

Jeon, E.-Y. and Day, R. R. (2016). The effectiveness of ER on reading proficiency: A meta-analysis. *Reading in a Foreign Language*, 28(2):246–265.

Jescheniak, J. D. and Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 20:824–843.

Johnson, R. C., Thompson, C. W., and Frincke, G. (1960). Word values, word frequency, and visual duration thresholds. *Psychological Review*, 67:332–342.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329–354.

Kincaid, J. P., Rogers, R. L., Fishburne, R. P., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index , fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Millington, Tennessee.

Kintsch, W., Britton, B. K., Fletcher, C. R., Kintsch, E., Mannes, S. M., and Nathan, M. J. (1993). A comprehension-based approach to learning and understanding. In Medin, D. L., editor, *Psychology of Learning and Motivation*, volume 30, pages 165–214. Academic Press, New York.

Kintsch, W. and Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In Nilsson, L. G., editor, *Perspectives on memory research*, pages 24–62. Erlbaum, Hillsdale, NJ.

Klare, G. R. (1968). The role of word frequency in readability. *Elementary English*, 45(1):12–22.

Klare, G. R., Mabry, J. E., and Gustafson, L. M. (1955). The relationship of style difficulty to immediate retention and to acceptability to technical material. *Journal of Educational Psychology*, 46:287–295.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing*, 2(1):76–104.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. Guilford, New York, 3rd edition.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, 20(2):148–161.

Kormos, J. and Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech ofsecond language learners. *System*, 32(2):145–164.

Krashen, S. (1985). *The Input Hypothesis: Issues and Implications*. Longman, New York.

Kretzenbacher, H. L. (1991). Syntax des wissenschaftlichen fachtextes. *Fachsprache. International Journal of LSP*, 4:118–137.

Krivanek, J. and Meurers, D. (2011). Comparing rule-based and data-driven dependency parsing of learnerlanguage. In *Proceedings of the Intern. Conference on Dependency Linguistics(DEPLING 2011)*, pages 310–317, Barcelona.

Kuiken, F. and Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics in Language Teaching*, 45(3):261–284.

Kuiken, F. and Vedder, I. (2008). Cognitive task complexity and written output in italian and French as a foreign language. *Journal of Second Language Writing*, 17(1):48–60.

Kusters, W. (2008). Complexity in linguistic theory, language learning and languagechange. In Miestamo, M., Sinnenmäki, K., and Karlsson, F., editors, *Language Complexity: Typology, contact, change*, pages 3–22. John Benjamins, Amsterdam.

Kučera, H. and Francis, W. N. (1967). *Computational Analysis of Present-day English*. Brown University Press, Providence, RI.

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. PhD thesis, Georgia State University.

Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

Lambert, C. and Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, 35(5):607–614.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, 12(4):439–448.

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

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In Bejoint, H. and Arnaud, P., editors, *Vocabulary and applied linguistics*, pages 126–132. Macmillan, Basingstoke & London.

Laufer, B. and Nation, P. (1995). Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, 16(3):307–322.

Laufer, B. and Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.

Lennon, C. and Burdick, H. (2014). The lexile framework as an approach for reading measurement and success. Technical report, MetaMetrics, Inc.

Lennon, P. (1990). Investigating fluency in efl: A quantitative approach. *Language Learning*, 40(3):387–417.

Leonard, K. R. and Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern Language Journal*, Advanced Online Access.

Leroy, G. and Kauchak, D. (2014). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*, 21:169–172.

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

Levelt, W. (1999). Producing spoken language: A blueprint of the speaker. In Brown, C. and Hagoort, P., editors, *The neurocognition of language*, pages 83–122. Oxford University Press.

Levy, R. and Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *The 5th International Conference on Language Resources and Evaluation*.

Lexile (2007). The Lexile Framework® for reading: Theoretical framework and development. Technical report, MetaMetrics, Inc., Durham, NC.

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

Lively, B. A. and Pressey, S. L. (1923). A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision*, 9:389–398.

Long, M. H. (1996). The role of linguistic environment in second language acquisition. In Ritchie, W. C. and Bhatia, T. K., editors, *Handbook of second language acquisition*, pages 413–468. Academic Press, New York.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., and Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. *Proceedings of the Cognitive Science Society*, 26(26).

Lowie, W. and Verspoor, M. (2015). Variability and variation in second language acquisition orders: A dynamic reevaluation. *Language Learning*, 65(1):63–88.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1):36–62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Lu, X. and Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29:16–27.

Lu, X., Gamson, D. a., and Eckert, S. A. (2014). Lexical difficulty and diversity of american elementary school reading textbooks: Changes over the past century. *International Journal of Corpus Linguistics*, 19(1):94–117.

Malvern, D. and Richards, B. (2002). Investigating accommodation in language prociency interviews using a new measure of lexical diversity. *Language Testing*, 19:85–104.

Malvern, D., Richards, B., Chipere, N., and Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan, New York.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Marks, C. B., Doctorow, M. J., and Wittrock, M. C. (1974). Word frequency and reading comprehension. *The Journal of Educational Research*, 67(6):259–262.

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, 2(4):260–293.

Mazgutova, D. and Kormos, J. (2015). Syntactic and lexical development in an intensive English for academic purposes programme. *Journal of Second Language Writing*, 29:3–15.

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

McClure, E. (1991). A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning*, 2:141–154.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, 8(3):299–325.

McDonough, K. and Mackey, A. (2006). Responses to recasts: Repetitions, primed production, and linguistic development. *Language Learning*, 56(4):693–720.

McDonough, K. and Mackey, A. (2008). syntactic priming and ESL question development. *Studies in Second Language Acquisition*, 30:31–47.

McLaughlin, G. H. (1969). Smog grading—a new readability formula. *Journal of Reading*, 22:639–646.

McNamara, D. A., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, Cambridge, M.A.

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2009). Linguistic features of writing quality. *Written Communication*, pages 1–30.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C. (2010). Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47:292–330.

Mesmer, H. A., Cunningham, J. W., and Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, 47(3):235–258.

Meunier, F. and Gouverneur, C. (2009). New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In Aijmer, K., editor, *Corpora and Language Teaching*, pages 179–201. John Benjamins, Amsterdam.

Meurers, D. (2012). Natural language processing and language learning. In Chapelle, C. A., editor, *Encyclopedia of Applied Linguistics*, pages 4193–4205. Wiley, Oxford.

Meurers, D. and Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(S1):66–95.

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., and Ott, N. (2010). Enhancing authentic web pages for language learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 10–18, Los Angeles. ACL.

Michaud, L. and McCoy, K. (2006). Capturing the evolution of grammatical knowledge in a CALL system for deaf learners of English. *International Journal of Artificial Intelligence in Education*, 16(1):65–97.

Michel, M. C., Kuiken, F., and Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45(3):241–259.

Miestami, M. (2008). Grammatical complexity in a cross-linguistic perspective. In Miestamo, M., Sinnenmäki, K., and Karlsson, F., editors, *Language Complexity: Typology, contact, change*, pages 23–41. John Benjamins, Amsterdam.

Milone, M. and Biemiller, A. (2014). The development of ATOS: The renaissance readability formula. Technical report, Renaissance Learning, Wisconsin Rapids.

Monsell, S., Doyle, M. C., and Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118:43–71.

Mora, J. C. (2006). Age effects on oral fluency development. In Muñoz, C., editor, *Age and the rate of foreign language learning*, pages 65–88. Multilingual Matters, Clevedon.

Murakami, A. (2013). *Individual Variation and the Role of L1 in the L2 Development of English Grammatical Morphemes: Insights From Learner Corpora*. PhD thesis, University of Cambridge.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of english grammatical morphemes. *Language Learning*, 6(4):834–871.

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La Revue canadienne des langues vivantes*, 63(1):59–82.

Neary-Sundquist, C. A. (2017). Syntactic complexity at multiple proficiency levels of L2 German speech. *International Journal of Applied Linguistics*, 27(1):242–262.

Nelson, J., Perfetti, C., Liben, D., and Liben, M. (2012). Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Council of Chief State School Officers.

Nobuyoshi, J. and Ellis, R. (1993). Focused communication tasks and second language acquisition. *ELT Journal*, 47(3):203–210.

Norrby, C. and Håkansson, G. (2007). The interaction of complexity and grammatical processability: The case of Swedish as a foreign language. *International Review of Applied Linguistics in Language Teaching*, 45:45–68.

Norris, J. and Ortega, L. (2003). Defining and measuring sla. In *The handbook of second language acquisition*, pages 716–761. Wiley Online Library.

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating caf in instructed sla: The case of complexity. *Applied Linguistics*, 30(4):555–578.

Ojemann, R. J. (1934). The reading ability of parents and factors associated with reading difficulty of parent education materials. *University of Iowa Studies in Child Welfare*, 8:11–32.

Ong, J. and Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, 19(4):218 – 233.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21(1):109–148.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners*. Unpublished doctoral dissertation, University of Hawaii, Manoa, HI.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college level L2 writing. *Applied Linguistics*, 24(4):492–518.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In Szmrecsanyi, B. and Kortmann, B., editors, *Linguistic complexity: Second language acquisition, indigenization, contact*, pages 127–155. de Gruyter, Berlin.

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, 29:82–94.

Ortega, L. and Sinicrope, C. (2008). Novice proficiency in a foreign language: A study of task-based performance profiling on the STAMP test. Technical report, Center for Applied Second Language Studies, University of Oregon.

Ott, N. and Ziai, R. (2010). Evaluating dependency parsing performance on German learner language. In Dickinson, M., Müürisep, K., and Passarotti, M., editors, *Proceedings of the Ninth International Workshop on Treebanks andLinguistic Theories*, volume 9 of *NEALT Proceeding Series*, pages 175–186, Tartu, Estonia. Tartu University Press.

Pallotti, G. (2008). Defining and assessing interactional complexity: An empirical study. In *AILA*, Essen.

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

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Panther, K.-U. (1981). Eine typisch indirekte sprachliche handlung imwissenschaftlichen diskurs. In Bungarten, T., editor, *Wissenschaftssprache: Beiträge zu Methode, Theorieund Despriktion*, pages 231–260. Wilhelm Fink Verlag.

Patty, W. W. and Painter, W. I. (1931). A technique for measuring the vocabulary burden of textbooks. *Journal of Educational Research*, 24:127–134.

Pearson, P. D. and Hiebert, E. H. (2014). The state of the field: Qualitative analyses of text complexity. *The Elementary School Journal*, 115(2):161–183.

Polat, B. and Kim, Y. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics*, 35(2):184–207.

Polio, C. and Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, 26:10–27.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review*, 56:282–308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52:513–536.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14:191–201.

Reid, J. (1992). A computer text analysis of four cohesion devices in english discourse by native and nonnative writers. *Journal of Second Language Writing*, 1(2):79–107.

Rescher, N. (1998). *Complexity: A philosophical overview*. Transaction Publishers, London.

Révész, A., Kourtali, N.-E., and Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, 67(1):208–241.

Reynolds, R. (2016). Insights from Russian second language readability classification: complexity-dependent training requirements, and featureevaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, California. Association for Computational Linguistics.

Reynolds, R., Schaf, E., and Meurers, D. (2014). A view of Russian: Visual input enhancement and adaptive feedback. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107, pages 98–112, Uppsala. ACL.

Rimrott, A. and Heift, T. (2008). Evaluating automatic detection of misspellings in german. *Language Learning and Technology*, 12(3):73–92.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, 45(2):283–331.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, 22(1):27–57.

Robinson, P. (2003). Attention and memory during sla. In *The handbook of second language acquisition*, pages 631–678. Blackwell Publishing Ltd.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 43(1):1–32.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In Robinson, P., editor, *Second language task complexity. Researching the cognition hypothesis of language learning and performance*, pages 3–37. John Benjamins.

Robinson, P. and Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching*, 45(3):161–176.

Rodrigo, V., Krashen, S., and Gribbons, B. (2004). The effectiveness of two comprehensible-input approaches to foreign language instruction at the intermediate level. *System*, 32(1):53–60.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Ryder, R. J. and Slater, W. H. (1988). The relationship between word frequency and word knowledge. *The Journal of Educational Research*, 81(5):312–317.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System*, 45:79–91.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In Schmidt, R., editor, *Attention and awareness in foreign language learning*, pages 1–63. University of Hawaii, Honolulu, HI.

Schmitt, N., Jiang, X., and Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(i):26–43.

Schmitt, N. and Redwood, S. (2011). Learner knowledge of phrasal verbs: A corpus-informed study. In Meunier, F., De Cock, S., Gilquin, G., and Paquot, M., editors, *A Taste for Corpora. In Honour of Sylviane Granger*, pages 173–207. John Benjamins Publishing Company, Amsterdam.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 99(6):323–338.

Schumacker, R. E. and Lomax, R. G. (2010). *A beginner's guide to structural equation modeling*. Routledge, New York, 3rd edition.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., and Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC) at COLING*, Osaka.

Shermis, M. D. and Burstein, J., editors (2013). *Handbook on Automated Essay Evaluation: Current Applications and New Directions*. Routledge, Taylor & Francis Group, London and New York.

Shin, J.-A. and Christianson, K. (2012). Structural priming and second language learning. *Language Learning*, 62(3):931–964.

Silva, T. (1992). L1 vs L2 writing: ESL graduate students' perceptions. *TESL Canada Journal*, 10(1):27–47.

Skalban, Y., Ha, L. A., Specia, L., and Mitkov, R. (2012). Automatic question generation in multimedia-based learning. In *Proceedings of COLING 2012: Posters*, pages 1151–1160, Mumbai, India. COLING.

Skehan, P. (1989). *Individual Differences in Second Language Learning*. Edward Arnold, London.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, 17(1):38.

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4):510–532.

Skehan, P. and Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1(3):185–211.

Sotillo, S. M. (2000). Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning & Technology*, 4(1):82–119.

Spada, N. and Tomita, Y. (2010). Interactions between type of instruction and type of languagefeature: A meta-analysis. *Language Learning*, 60(2):263–308.

Spoelman, M. and Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of finnish. *Applied Linguistics*, 31(4):532.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, 18(2):103–118.

Sturgis, P. (2016). Structural Equation Modeling: What is it and what can we use it for? [accessed: September 22, 2017].

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In Gass, S. M. and Madden, C. G., editors, *Input in second language acquisition*, pages 235–253. Newbury House, Rowley, MA.

Tabari, M. A. (2016). The effects of planning time on complexity, accuracy, fluency, and lexical variety in L2 descriptive writing. *Asian-Pacific Journal of Second and Foreign Language Education*, 1(1).

Taguchi, N., Crawford, W., and Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? a case of argumentative essays in a college composition program. *TESOL Quarterly*, 47(2):420–430.

Taylor, B. P. (1975). The use of overgeneralization and transfer learning strategies by elementary and intermediate students of ESL. *Language Learning*, 25(1):73–107.

Taylor, Annand Marcus, M. S. B. (2003). The penn treebank: An overview. In Abeillé, A., editor, *Treebanks: Building and Using Parsed Corpora*, pages 5–22. Springer Netherlands, Dordrecht.

Thorndike, E. (1921). *The Teacher's Word Book*. Teachers College, Columbia University, New York.

Tono, Y. (2004). Multiple comparisons of IL, L1, and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learers of English. In Aston, G., Bernardini, S., and Stewart, D., editors, *Corpora and Language Learners*. John Benjamins.

Towell, R. (2012). Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 47–69. John Benjamins.

Towell, R. and Hawkins, R. (1994). *Approaches to second language acquisition*. Multilingual Matters, Clevedon.

Tracy-Ventura, N. and Myles, F. (2015). The importance of task variability in the design of learner corpora for sla research. *International Journal of Learner Corpus Research*, 1(1):58–95.

Ulijn, J. M. and Strother, J. B. (1990). The effect of syntactic simplification on reading est texts as l1 and l2. *Journal of Research in Reading'*, 13:38–54.

Vajjala, S. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. PhD thesis, University of Tübingen.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, Montreal, Canada. Association of Computational Linguistics.

Vajjala, S. and Meurers, D. (2014). Readability assessment for text simplification: From analyzingdocuments to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue onCurrent Research in Readability and Text Simplification*, 165(2):142–222.

van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). Subtlex-uk: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–90.

VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, 12(3):287–301.

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 38(1):90–111.

Verspoor, M. and Dijk, M. V. (2012). Variability in a dynamic systems theory approach to second languageacquisition. In Chapelle, C. A., editor, *The Encyclopedia of Applied Linguistics*, pages 6051–6059. Blackwell Publishing, Oxford.

Verspoor, M., Lowie, W., and Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal*, 92(2):214–231.

Verspoor, M., Schmid, M. S., and Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21(3):239–263.

Vogel, M. and Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *Elementary School Journal*, 28:373–381.

Vor der Brück, T. and Hartrumpf, S. (2007). A semantically oriented readability checker for German. In Vetulani, Z., editor, *Proceedings of the 3rd Language & Technology Conference*, pages 270–274, Poznań, Poland. Wydawnictwo Poznańskie.

vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008). A readability checker with supervised learning using deep syntactic and semantic indicators. In Erjavec, T.

and Gros, J. v., editors, *Proceedings of the 11th International Multiconference: Information Society - IS 2008 - Language Technologies*.

Vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008). A readability checker with supervised learning using deep syntacticand semantic indicators. *Informatica*, 32(4):429–435.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, 96(4):576–598.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal*, 97(S1):11–30.

Vyatkina, N. (2015). New developments in the study of L2 writing complexity: An editorial. *Journal of Second Language Writing*, 29:1 – 2.

Vyatkina, N., Hirschmann, H., and Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing*, 29:28–50.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

Wang, C. and Wang, M. (2015). Effect of alignment on L2 written production. *Applied Linguistics*, 36(5):503–526.

Way, D. P., Joiner, E. G., and Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal*, 84(2):171–184.

Weiß, Z. (2017). Using measures of linguistic complexity to assess german l2 proficiency in learner corpora under consideration of task-effects. Master's thesis, Universitä Tübingen, Germany.

Wendel, J. N. (1997). *Planning and Second-language Narrative Production*. Unpublished doctoral dissertation, Temple University, Temple.

White, L. (1990). Second language acquisition and universal grammar. *Studies in Second Language Acquisition*, 12:121–133.

White, L. (1991). The verb movement parameter in second language acquisition. *Language Acquisition*, 1:337–360.

White, L. (1992). On triggering data in L2 acquisition: A reply to Schwartz and Gubala-Ryzak. *Second Language Research*, 8:120–137.

White, L. (2003). *Second language acquisition and universal grammar*. Cambridge University Press, Cambridge.

Williams, S., Siddharthan, A., and Nenkova, A., editors (1994). *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Gothenburg, Sweden. ACL.

Wisniewski, K. (2017). The empirical validity of the common european framework of reference scales. an exemplary study for the vocabulary and fluency scales in a language testing context. *Applied Linguistics*, Advanced Online Access.

Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy, & complexity. Technical report, Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Yang, W., Lu, X., and Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28:53–67.

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, 66:130–141.

Yoon, H.-J. and Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, 51(2):257–301.

Yuan, F. and Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24(1):1–27.

Zakaluk, B. L. and Samuels, S. J., editors (1988). *Readability: Its Past, Present, and Future*. International Reading Association, Newark, Del.

Zhong, H. F. (2016). The relationship between receptive and productive vocabulary knowledge: a perspective from vocabulary use in sentence writing. *The Language Learning Journal*, Advanced Access.

Appendix A

List of complexity measures

This appendix lists the full set of complexity measures used in a few studies in this dissertation. The integration of the full measure set into the CTAP platform is still underway, but the first relase of the system has included a significant subset of the following list.

Index	Measure
Lexical Density Measures (1–100)	
1	Number of adjective lemmas
2	Number of adverb lemmas
3	Number of all word tokens, excluding number and punctuation tokens
4	Number of adjective types
5	Number of adverb types
6	Number of lexical tokens
7	Number of noun types
8	Number of verb types
9	Number of all tokens, including punctuations and numbers.
10	Number of coordinating conjunction tokens.
11	Number of coordinating conjunction types.
12	Number of cardinal number tokens.
13	Number of cardinal number types.
14	Number of determiner tokens.
15	Number of determiner types.
16	Number of existential there tokens.
17	Number of existential there Types.
18	Number of foreign word tokens.

Index	Measure
19	Number of foreign word Types.
20	Number of preposition and subordinating conjunction tokens.
21	Number of preposition and subordinating conjunction types.
22	Number of adjective tokens.
23	Number of adjective types.
24	Number of comparative adjective tokens.
25	Number of comparative adjective types.
26	Number of superlative adjective tokens.
27	Number of superlative adjective types.
28	Number of list item tokens.
29	Number of list item types.
30	Number of modal word tokens.
31	Number of modal word types.
32	Number of singular or mass noun tokens.
33	Number of singular or mass noun types.
34	Number of plural noun tokens.
35	Number of plural noun types.
36	Number of singular proper noun tokens.
37	Number of singular proper noun types.
38	Number of plural proper noun tokens.
39	Number of plural proper noun types.
40	Number of predeterminer tokens.
41	Number of predeterminer types.
42	Number of possessive ending tokens.
43	Number of possessive ending types.
44	Number of personal pronoun tokens.
45	Number of personal pronoun types.
46	Number of possessive pronoun tokens.
47	Number of lexical lemmas
48	Number of noun lemmas
49	Number of verb lemmas
50	Number of word types, excluding number and punctuation types
51	Number of numeric tokens
52	Number of nuique number tokens
53	Number of possessive pronoun types.
54	Number of adverb tokens.

Index	Measure
55	Number of adverb types.
56	Number of comparative adverb tokens.
57	Number of comparative adverb types.
58	Number of superlative adverb tokens.
59	Number of superlative adverb types.
60	Number of particle tokens.
61	Number of particle types.
62	Number of symbol tokens.
63	Number of symbol types.
64	Number of ‘to’ tokens.
65	Number of ‘to’ types.
66	Number of interjection tokens.
67	Number of interjection types.
68	Number of verb tokens in their base form
69	Number of verb types in their base form
70	Number of verb tokens in their past form
71	Number of verb types in their past form
72	Number of verb tokens in their gerund or present participle form
73	Number of verb types in their gerund or present participle form
74	Number of verb tokens in their past participle form
75	Number of verb types in their past participle form
76	Number of verb tokens in their non-third person singular present form
77	Number of verb types in their non-third person singular present form
78	Number of verb tokens in their third person singular present form
79	Number of verb types in their third person singular present form
80	Number of wh-determiner tokens
81	Number of determiner types
82	Number of wh-pronoun tokens
83	Number of wh-pronoun types
84	Number of possessive wh-pronoun tokens
85	Number of possessive wh-pronoun types
86	Number of wh-adverb tokens
87	Number of wh-adverb types
88	Number of punctuation mark tokens

Index	Measure
89	Number of all word types
90	Number of punctuation mark types
91	Number of words that are used only once
92	Number of syllables
93	Number of adjective tokens
94	Number of adverb tokens
95	Number of noun tokens
96	Number of verb tokens
97	Number of lexical types
98	Ratio of modal tokens to all word tokens
99	Ratio of modal tokens to all verb tokens, including modal verbs
100	Ratio of lexical tokens to all word tokens
Lexical Sophistication Measures (101–460)	
101	Mean token frequency of adjectives calculated with the BNC frequency list
102	Standard deviation of token frequency of adjectives calculated with the BNC frequency list
103	Mean token frequency of adverbs calculated with the BNC frequency list
104	Standard deviation of token frequency of adverbs calculated with the BNC frequency list
105	Mean token frequency of all words calculated with the BNC frequency list
106	Standard deviation of token frequency of all words calculated with the BNC frequency list
107	Mean token frequency of all lexical words calculated with the BNC frequency list
108	Standard deviation of token frequency of all lexical words calculated with the BNC frequency list
109	Mean token frequency of all lexical words calculated with the BNC frequency list
110	Standard deviation of token frequency of all nouns calculated with the BNC frequency list
111	Mean token frequency of all verbs calculated with the BNC frequency list

Index	Measure
112	Standard deviation of token frequency of all verbs calculated with the BNC frequency list
113	Mean type frequency of all adjectives calculated with the BNC frequency list
114	Standard deviation of type frequency of all adjectives calculated with the BNC frequency list
115	Mean type frequency of all advers calculated with the BNC frequency list
116	Standard deviation of type frequency of all adverbs calculated with the BNC frequency list
117	Mean type frequency of all words calculated with the BNC frequency list
118	Standard deviation of type frequency of all words calculated with the BNC frequency list
119	Mean type frequency of all lexical words calculated with the BNC frequency list
120	Standard deviation of type frequency of lexical words calculated with the BNC frequency list
121	Mean type frequency of all nouns calculated with the BNC frequency list
122	Standard deviation of type frequency of nouns calculated with the BNC frequency list
123	Mean type frequency of all verbs calculated with the BNC frequency list
124	Standard deviation of type frequency of verbs calculated with the BNC frequency list
125	Mean token frequency of adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure
126	SD of token frequency of adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure
127	Mean token frequency of adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure
128	SD of token frequency of adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure
129	Mean token frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure

Index	Measure
130	SD of token frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure
131	Mean token frequency of all lexical words calculated with the SUBTLEXus frequency list's Contextual Diversity measure
132	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's Contextual Diversity measure
133	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure
134	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure
135	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure
136	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure
137	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure
138	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure
139	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure
140	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure
141	Mean type frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure
142	SD of type frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure
143	Mean type frequency of all lexical words calculated with the SUBTLEXus frequency list's Contextual Diversity measure
144	SD of type frequency of all lexicals calculated with the SUBTLEXus frequency list's Contextual Diversity measure
145	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure
146	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure
147	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure

Index	Measure
148	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure
149	Mean token frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 WF measure
150	SD of token frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 WF measure
151	Mean token frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure
152	SD of token frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure
153	Mean token frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure
154	SD of token frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure
155	Mean token frequency of all lexicals calculated with the SUBTLEXus frequency list's Log10 WF measure
156	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 WF measure
157	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure
158	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure
159	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure
160	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure
161	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 WF measure
162	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 WF measure
163	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure
164	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure
165	Mean type frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure

Index	Measure
166	SD of type frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure
167	Mean type frequency of all lexicals calculated with the SUBTLEXus frequency list's Log10 WF measure
168	SD of type frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 WF measure
169	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure
170	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure
171	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure
172	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure
173	CTTR of adjectives that are not in the list of the top 2000 most frequent words of the New General Service List
174	CTTR of adverbs that are not in the list of the top 2000 most frequent words of the New General Service List
175	CTTR of all words that are not in the list of the top 2000 most frequent words of the New General Service List
176	CTTR of all lexical words that are not in the list of the top 2000 most frequent words of the New General Service List
177	CTTR of all nouns that are not in the list of the top 2000 most frequent words of the New General Service List
178	CTTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List
179	GTTR of all adjectives that are not in the list of the top 2000 most frequent words of the New General Service List
180	GTTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List
181	GTTR of all words that are not in the list of the top 2000 most frequent words of the New General Service List
182	GTTR of all lexical words that are not in the list of the top 2000 most frequent words of the New General Service List
183	GTTR of all nouns that are not in the list of the top 2000 most frequent words of the New General Service List

Index	Measure
184	GTTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List
185	LogTTR of all adjectives that are not in the list of the top 2000 most frequent words of the New General Service List
186	LogTTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List
187	LogTTR of all words that are not in the list of the top 2000 most frequent words of the New General Service List
188	LogTTR of all lexical words that are not in the list of the top 2000 most frequent words of the New General Service List
189	LogTTR of all nouns that are not in the list of the top 2000 most frequent words of the New General Service List
190	LogTTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List
191	Ratio of sophisticated adjective tokens, which are words that are not in the top 2000 most frequent words from NGSL, to all adjective tokens
192	Number of adjective tokens that are not in the list of the top 2000 most frequent words of the New General Service List
193	Ratio of sophisticated adverb tokens, which are words that are not in the top 2000 most frequent words from NGSL, to all adverb tokens
194	Number of adverbs tokens that are not in the list of the top 2000 most frequent words of the New General Service List
195	Ratio of sophisticated word tokens, which are words that are not in the top 2000 most frequent words from NGSL, to all word tokens
196	Number of all word tokens that are not in the list of the top 2000 most frequent words of the New General Service List
197	Ratio of sophisticated lexical word tokens, which are words that are not in the top 2000 most frequent words from NGSL, to all lexical word tokens
198	Number of lexical tokens that are not in the list of the top 2000 most frequent words of the New General Service List

Index	Measure
199	Ratio of sophisticated noun tokens, which are words that are not in the top 2000 most frequent words from NGSL, to all noun tokens
200	Number of noun tokens that are not in the list of the top 2000 most frequent words of the New General Service List
201	Ratio of sophisticated verbs tokens, which are words that are not in the top 2000 most frequent words from NGSL, to all verbs tokens
202	Number of verb tokens that are not in the list of the top 2000 most frequent words of the New General Service List
203	Ratio of sophisticated adjective types, which are words that are not in the top 2000 most frequent words from NGSL, to all adjective types
204	Number of adjective types that are not in the list of the top 2000 most frequent words of the New General Service List
205	Ratio of sophisticated adverb types, which are words that are not in the top 2000 most frequent words from NGSL, to all adverb types
206	Number of adverb types that are not in the list of the top 2000 most frequent words of the New General Service List
207	Ratio of sophisticated word types, which are words that are not in the top 2000 most frequent words from NGSL, to all word types
208	Number of all word types that are not in the list of the top 2000 most frequent words of the New General Service List
209	Ratio of sophisticated lexical types, which are words that are not in the top 2000 most frequent words from NGSL, to all lexical types
210	Number of lexical types that are not in the list of the top 2000 most frequent words of the New General Service List
211	Ratio of sophisticated noun types, which are words that are not in the top 2000 most frequent words from NGSL, to all noun types
212	Number of noun types that are not in the list of the top 2000 most frequent words of the New General Service List
213	Ratio of sophisticated verb types, which are words that are not in the top 2000 most frequent words from NGSL, to all verb types

Index	Measure
214	Number of verb types that are not in the list of the top 2000 most frequent words of the New General Service List
215	STTR of all adjectives that are not in the list of the top 2000 most frequent words of the New General Service List
216	STTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List
217	STTR of all words that are not in the list of the top 2000 most frequent words of the New General Service List
218	STTR of all lexical words that are not in the list of the top 2000 most frequent words of the New General Service List
219	STTR of all nouns that are not in the list of the top 2000 most frequent words of the New General Service List
220	STTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List
221	TTR of all adjectives that are not in the list of the top 2000 most frequent words of the New General Service List
222	TTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List
223	TTR of all words that are not in the list of the top 2000 most frequent words of the New General Service List
224	TTR of all lexicals that are not in the list of the top 2000 most frequent words of the New General Service List
225	TTR of all nouns that are not in the list of the top 2000 most frequent words of the New General Service List
226	TTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List
227	UberTTR of all adjectives that are not in the list of the top 2000 most frequent words of the New General Service List
228	UberTTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List
229	UberTTR of all words that are not in the list of the top 2000 most frequent words of the New General Service List
230	UberTTR of all lexical words that are not in the list of the top 2000 most frequent words of the New General Service List
231	UberTTR of all nouns that are not in the list of the top 2000 most frequent words of the New General Service List

Index	Measure
232	UberTTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List
233	Mean token frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure
234	SD of token frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure
235	Mean token frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure
236	SD of token frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure
237	Mean token frequency of all words calculated with the SUBTLEXus frequency list's WF measure
238	SD of token frequency of all words calculated with the SUBTLEXus frequency list's WF measure
239	Mean token frequency of all lexical words calculated with the SUBTLEXus frequency list's WF measure
240	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's WF measure
241	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure
242	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure
243	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure
244	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure
245	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure
246	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure
247	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure
248	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure
249	Mean type frequency of all words calculated with the SUBTLEXus frequency list's WF measure

Index	Measure
250	SD of type frequency of all words calculated with the SUBTLEXus frequency list's WF measure
251	Mean type frequency of all lexicals calculated with the SUBTLEXus frequency list's WF measure
252	SD of type frequency of all lexical words calculated with the SUBTLEXus frequency list's WF measure
253	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure
254	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure
255	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure
256	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure
257	Mean token frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure
258	SD of token frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure
259	Mean token frequency of adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure
260	SD of token frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure
261	Mean token frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure
262	SD of token frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure
263	Mean token frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure
264	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure
265	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure
266	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure
267	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure

Index	Measure
268	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure
269	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure
270	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure
271	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure
272	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure
273	Mean type frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure
274	SD of type frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure
275	Mean type frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure
276	SD of type frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure
277	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure
278	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure
279	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure
280	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure
281	CTTR of adjectives that are not in the list of the top 2000 most frequent words of BNC
282	CTTR of adverbs that are not in the list of the top 2000 most frequent words of BNC
283	CTTR of all words that are not in the list of the top 2000 most frequent words of BNC
284	CTTR of all lexicals that are not in the list of the top 2000 most frequent words of BNC
285	CTTR of nouns that are not in the list of the top 2000 most frequent words of BNC

Index	Measure
286	CTTR of verbs that are not in the list of the top 2000 most frequent words of BNC
287	GTTR of adjectives that are not in the list of the top 2000 most frequent words of BNC
288	GTTR of adverbs that are not in the list of the top 2000 most frequent words of BNC
289	GTTR of all words that are not in the list of the top 2000 most frequent words of BNC
290	GTTR of all lexical words that are not in the list of the top 2000 most frequent words of BNC
291	GTTR of nouns that are not in the list of the top 2000 most frequent words of BNC
292	GTTR of verbs that are not in the list of the top 2000 most frequent words of BNC
293	LogTTR of adjectives that are not in the list of the top 2000 most frequent words of BNC
294	LogTTR of adverbs that are not in the list of the top 2000 most frequent words of BNC
295	LogTTR of all words that are not in the list of the top 2000 most frequent words of BNC
296	LogTTR of all lexical words that are not in the list of the top 2000 most frequent words of BNC
297	LogTTR of nouns that are not in the list of the top 2000 most frequent words of BNC
298	LogTTR of verbs that are not in the list of the top 2000 most frequent words of BNC
299	Ratio of sophisticated adjective tokens, which are words that are not in the top 2000 most frequent words from BNC, to all adjective tokens
300	Number of sophisticated adjectives tokens that are not in the list of the top 2000 most frequent words of BNC
301	Ratio of sophisticated adverb tokens, which are words that are not in the top 2000 most frequent words from BNC, to all adverb tokens
302	Number of sophisticated adverb tokens that are not in the list of the top 2000 most frequent words of BNC

Index	Measure
303	Ratio of sophisticated tokens, which are words that are not in the top 2000 most frequent words from BNC, to all tokens
304	Number of sophisticated word tokens that are not in the list of the top 2000 most frequent words of BNC
305	Ratio of sophisticated lexical tokens, which are words that are not in the top 2000 most frequent words from BNC, to all lexical tokens
306	Number of sophisticated lexical tokens that are not in the list of the top 2000 most frequent words of BNC
307	Ratio of sophisticated noun tokens, which are words that are not in the top 2000 most frequent words from BNC, to all noun tokens
308	Number of sophisticated noun tokens that are not in the list of the top 2000 most frequent words of BNC
309	Ratio of sophisticated verb tokens, which are words that are not in the top 2000 most frequent words from BNC, to all verb tokens
310	Number of sophisticated verb tokens that are not in the list of the top 2000 most frequent words of BNC
311	Ratio of sophisticated adjective types, which are words that are not in the top 2000 most frequent words from BNC, to all adjective types
312	Number of sophisticated adjectives types that are not in the list of the top 2000 most frequent words of BNC
313	Ratio of sophisticated adverb types, which are words that are not in the top 2000 most frequent words from BNC, to all adverb types
314	Number of sophisticated adverb types that are not in the list of the top 2000 most frequent words of BNC
315	Ratio of sophisticated word types, which are words that are not in the top 2000 most frequent words from BNC, to all word types
316	Number of sophisticated word types that are not in the list of the top 2000 most frequent words of BNC
317	Ratio of sophisticated lexical types, which are words that are not in the top 2000 most frequent words from BNC, to all lexical types
318	Number of sophisticated lexical types that are not in the list of the top 2000 most frequent words of BNC

Index	Measure
319	Ratio of sophisticated noun types, which are words that are not in the top 2000 most frequent words from BNC, to all noun types
320	Number of sophisticated noun types that are not in the list of the top 2000 most frequent words of BNC
321	Ratio of sophisticated verb types, which are words that are not in the top 2000 most frequent words from BNC, to all verb types
322	Number of sophisticated verb types that are not in the list of the top 2000 most frequent words of BNC
323	STTR of all adjectives that are not in the list of the top 2000 most frequent words of BNC
324	STTR of all adverbs that are not in the list of the top 2000 most frequent words of BNC
325	STTR of all words that are not in the list of the top 2000 most frequent words of BNC
326	STTR of all lexical words that are not in the list of the top 2000 most frequent words of BNC
327	STTR of all nouns that are not in the list of the top 2000 most frequent words of BNC
328	STTR of all verbs that are not in the list of the top 2000 most frequent words of BNC
329	TTR of all adjectives that are not in the list of the top 2000 most frequent words of BNC
330	TTR of all adverbs that are not in the list of the top 2000 most frequent words of BNC
331	TTR of all words that are not in the list of the top 2000 most frequent words of BNC
332	TTR of all lexical words that are not in the list of the top 2000 most frequent words of BNC
333	TTR of all nouns that are not in the list of the top 2000 most frequent words of BNC
334	TTR of all verbs that are not in the list of the top 2000 most frequent words of BNC
335	UberTTR of all adjectives that are not in the list of the top 2000 most frequent words of BNC
336	UberTTR of all adverbs that are not in the list of the top 2000 most frequent words of BNC

Index	Measure
337	UberTTR of all words that are not in the list of the top 2000 most frequent words of BNC
338	UberTTR of all lexical words that are not in the list of the top 2000 most frequent words of BNC
339	UberTTR of all nouns that are not in the list of the top 2000 most frequent words of BNC
340	UberTTR of all verbs that are not in the list of the top 2000 most frequent words of BNC
341	CTTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List
342	CTTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
343	CTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List
344	CTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List
345	CTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List
346	CTTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
347	GTTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List
348	GTTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
349	GTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List
350	GTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List
351	GTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List
352	GTTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
353	LogTTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List
354	LogTTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List

Index	Measure
355	LogTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List
356	LogTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List
357	LogTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List
358	LogTTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
359	Ratio of easy adjective tokens to all adjective tokens with NGSL top 1000 most frequent words as easy word list
360	Number of adjective tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List
361	Ratio of easy adverb tokens to all adverb tokens with NGSL top 1000 most frequent words as easy word list
362	Number of adverb tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List
363	Ratio of all easy tokens to all tokens with NGSL top 1000 most frequent words as easy word list
364	Number of all word tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List
365	Ratio of easy lexical tokens to all lexical tokens with NGSL top 1000 most frequent words as easy word list
366	Number of all lexical tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List
367	Ratio of easy noun tokens to all noun tokens with NGSL top 1000 most frequent words as easy word list
368	Number of noun tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List
369	Ratio of easy verb tokens to all verb tokens with NGSL top 1000 most frequent words as easy word list
370	Number of verb tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List
371	Ratio of easy adjective types to all adjective types with NGSL top 1000 most frequent words as easy word list
372	Number of adjective types that are in the list of easy words (top 1000 most frequent) from the New General Service List

Index	Measure
373	Ratio of easy adverb types to all adverb types with NGSL top 1000 most frequent words as easy word list
374	Number of adverb types that are in the list of easy words (top 1000 most frequent) from the New General Service List
375	Ratio of easy word types to all word types with NGSL top 1000 most frequent words as easy word list
376	Number of all word types that are in the list of easy words (top 1000 most frequent) from the New General Service List
377	Ratio of easy lexical types to all lexical types with NGSL top 1000 most frequent words as easy word list
378	Number of all lexical types that are in the list of easy words (top 1000 most frequent) from the New General Service List
379	Ratio of easy noun types to all noun types with NGSL top 1000 most frequent words as easy word list
380	Number of all noun types that are in the list of easy words (top 1000 most frequent) from the New General Service List
381	Ratio of easy verb types to all verb types with NGSL top 1000 most frequent words as easy word list
382	Number of verb types that are in the list of easy words (top 1000 most frequent) from the New General Service List
383	STTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List
384	STTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
385	STTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List
386	STTR of lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List
387	STTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List
388	STTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
389	TTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List
390	TTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List

Index	Measure
391	TTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List
392	TTR of lexicals that are in the list of easy words (top 1000 most frequent) from the New General Service List
393	TTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List
394	TTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
395	UberTTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List
396	UberTTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
397	UberTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List
398	UberTTR of lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List
399	UberTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List
400	UberTTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List
401	CTTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC
402	CTTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC
403	CTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC
404	CTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC
405	CTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC
406	CTTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC
407	GTTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC
408	GTTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC

Index	Measure
409	GTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC
410	GTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC
411	GTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC
412	GTTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC
413	LogTTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC
414	LogTTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC
415	LogTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC
416	LogTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC
417	LogTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC
418	LogTTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC
419	Ratio of easy adjective tokens to all adjective tokens with BNC top 1000 most frequent words as easy word list
420	Number of easy adjective lemma tokens from the BNC list of easy words (top 1000 most frequent)
421	Ratio of easy adverb tokens to all adverb tokens with BNC top 1000 most frequent words as easy word list
422	Number of easy adverb lemma tokens from the BNC list of easy words (top 1000 most frequent)
423	Ratio of easy tokens to all word tokens with BNC top 1000 most frequent words as easy word list
424	Number of easy lemma tokens from the BNC list of easy words (top 1000 most frequent)
425	Ratio of easy lexical tokens to all lexical tokens with BNC top 1000 most frequent words as easy word list
426	Number of easy lexical lemma tokens from the BNC list of easy words (top 1000 most frequent)

Index	Measure
427	Ratio of easy noun tokens to all noun tokens with BNC top 1000 most frequent words as easy word list
428	Number of easy noun lemma tokens from the BNC list of easy words (top 1000 most frequent)
429	Ratio of easy verb tokens to all verb tokens with BNC top 1000 most frequent words as easy word list
430	Number of easy verb lemma tokens from the BNC list of easy words (top 1000 most frequent)
431	Ratio of easy adjective types to all adjective types with BNC top 1000 most frequent words as easy word list
432	Number of easy adjective lemma types from the BNC list of easy words (top 1000 most frequent)
433	Ratio of easy adverb types to all adverb types with BNC top 1000 most frequent words as easy word list
434	Number of easy adverb lemma types from the BNC list of easy words (top 1000 most frequent)
435	Ratio of easy word types to all word types with BNC top 1000 most frequent words as easy word list
436	Number of easy lemma types from the BNC list of easy words (top 1000 most frequent)
437	Ratio of easy lexical types to all lexical types with BNC top 1000 most frequent words as easy word list
438	Number of easy lexical lemma types from the BNC list of easy words (top 1000 most frequent)
439	Ratio of easy noun types to all noun types with BNC top 1000 most frequent words as easy word list
440	Number of easy noun lemma types from the BNC list of easy words (top 1000 most frequent)
441	Ratio of easy verb types to all verb types with BNC top 1000 most frequent words as easy word list
442	Number of easy verb lemma types from the BNC list of easy words (top 1000 most frequent)
443	STTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC
444	STTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC

Index	Measure
445	STTR of all words that are in the list of easy words (top 1000 most frequent) from BNC
446	STTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC
447	STTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC
448	STTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC
449	TTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC
450	TTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC
451	TTR of all words that are in the list of easy words (top 1000 most frequent) from BNC
452	TTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC
453	TTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC
454	TTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC
455	UberTTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC
456	UberTTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC
457	UberTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC
458	UberTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC
459	UberTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC
460	UberTTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC
Lexical Sophistication Measures (461–505)	
461	Corrected type-token ratio for adjectives
462	Corrected type-token ratio for adverbs
463	Corrected type-token ratio for all words

Index	Measure
464	Corrected type-token ratio for all lexical words
465	Corrected type-token ratio for all nouns
466	Corrected type-token ratio for all verbs
467	Guiraud's type-token ratio for all adjectives
468	Guiraud's type-token ratio for all adverbs
469	Guiraud's type-token ratio for all words
470	Guiraud's type-token ratio for all lexicals
471	Guiraud's type-token ratio for all nouns
472	Guiraud's type-token ratio for all verbs
473	Bilogarithmic type-token ratio for all adjectives
474	Bilogarithmic type-token ratio for all adverbs
475	Bilogarithmic type-token ratio for all words
476	Bilogarithmic type-token ratio for all lexicals
477	Bilogarithmic type-token ratio for all nouns
478	Bilogarithmic type-token ratio for all verbs
479	Evenly segmented type-token ratio for all words, 10 segments
480	Ratio of number of adverb and adjective types to number of all lexical tokens
481	Mean segmented type-token ratio of all 50-word segments
482	Normalized type-token ratio for adjetives
483	Normalized type-token ratio for adverbs
484	Normalized type-token ratio for all words
485	Normalized type-token ratio for all lexical words
486	Normalized type-token ratio for all nouns
487	Normalized type-token ratio for all verbs
488	Ratio of adjective types to lexical tokens
489	Ratio of adverb types to lexical tokens
490	Ratio of noun types to lexical tokens
491	Ratio of verb types to lexical tokens
492	Type-token ratio of all adjectives
493	Type-token ratio of all adverbs
494	Type-token ratio of all words
495	Type-token ratio of all lexical words
496	Type-token ratio of all nouns
497	Type-token ratio of all verbs
498	Log-transformed type-token ratio of all adjectives

Index	Measure
499	Log-transformed type-token ratio of all adverbs
500	Log-transformed type-token ratio of all words
501	Log-transformed type-token ratio of all lexical words
502	Log-transformed type-token ratio of all nouns
503	Log-transformed type-token ratio of all verbs
504	A measure of the mean length of word strings that maintain a criterion level of lexical variation.
505	An index that derives from the hypergeometric distribution function
Syntactic Density Measures (506–535)	
506	Number of clauses
507	Number of complex norminals
508	Number of coordinate phrases
509	Number of T-units
510	Number of dependent clauses
511	Number of fragment clauses
512	Number of fragment T-units
513	Number of adjective clauses
514	Number of coordinate clauses
515	Number of nominal clauses
516	Number of sentences
517	Number of T-units
518	Number of verb phrases
519	Number of passive clasuses
520	Number of passive sentences
521	Number of adverbial clasuses
522	Number of adjective phrases
523	Number of adverb phrases
524	Number of noun phrases
525	Number of prepositional phrases
526	Number of declarative clauses
527	Number of subordinate clauses
528	Number of direct questions
529	Number of inverted declarative sentences
530	Number of yes/no questions
531	Number of Wh-phrases

Index	Measure
532	Number of Wh noun phrases
533	Number of Wh prepositional phrases
534	Number of Wh adjective phrases
535	Number of unknown constituents
Syntactic Index Measures (536–549)	
536	Mean global edit distance of parse tree with lemma
537	SD of global edit distance of parse tree with lemma
538	Mean local edit distance of parse tree with lemma
539	SD of local edit distance of parse tree with lemma
540	Mean global edit distance of parse tree with part-of-speech of words
541	SD of global edit distance of parse tree with part-of-speech of words
542	Mean local edit distance of parse tree with part-of-speech of words
543	SD of local edit distance of parse tree with part-of-speech of words
544	Mean of global edit distance of parse tree with word tokens
545	SD of global edit distance of parse tree with word tokens
546	Mean of local edit distance of parse tree with word tokens
547	SD of local edit distance of parse tree with word tokens
548	Mean of left embeddedness of all sentences
549	SD of left embeddedness of all sentences
Syntactic Ratio Measures (550–566)	
550	number of coordinate clauses per clause
551	number of complex nominals per clause
552	number of complex nominals per T-unit
553	number of complex nominals per noun phrase
554	number of clauses per sentence
555	number of clauses per T-unit
556	number of coordinate phrases per clause
557	number of coordinate phrases per T-unit
558	number of complex T-units per T-unit
559	number of dependent clauses per clause
560	number of dependent clauses per T-unit
561	Mean length of clause in tokens
562	Mean length of sentence in tokens
563	Mean length of sentence in syllables

Index	Measure
564	Mean length of T-units in tokens
565	number of T-units per sentence
566	number of verb phrases per T-unit
Discourse Cohesion Measures (567–576)	
567	Global argument (nouns and pronouns) overlap
568	Mean global lexical overlap
569	SD of global lexical overlap
570	Global noun overlap
571	Global stem (nouns, pronouns, verbs, adjectives, and adverbs) overlap
572	Local argument overlap
573	Mean local lexical overlap
574	SD of local lexical overlap
575	Local noun overlap
576	Local stem overlap

Table A.1: Full list of the comprehensive set of complexity measures used in this dissertation

Appendix B

List of accuracy measures

Table B.1 lists the measures used in the study on trade-off between complexity and accuracy reported in Chapter 4. These accuracy measures were calculated from the manual annotation of Göpferich and Neumann (2016).

Index	Measure
1	#errors / 100 words
2	% error-free sentences
3	# aspect errors
4	# spelling errors
5	# blending errors
6	# case/number/agreement errors
7	# collocation errors
8	# cultural specificity errors
9	# formatting errors
10	# functional sentence perspective errors (FSP)
11	# idiomaticity/genre errors
12	# implicitness errors
13	# infinitive/participle errors
14	# modality/illocution errors
15	# mood errors
16	# other grammar errors
17	# preposition errors
18	# punctuation errors
19	# redundancy errors
20	# repetition errors
21	# rhetoric errors

Index	Measure
22	# secondary subjectivization errors
23	# semantic errors
24	# sense errors
25	# specifier (article or determiner) errors
26	# syntax errors
27	# tense errors
28	# text coherence errors
29	# valency errors
30	# voice errors
31	# word form errors

Table B.1: A complete list of accuracy measures used in the study reported in Chapter 4

Appendix C

Statistics of regression models from the vector distance study

Linear regression models were fitted by regressing *improvement* on *challenge*. For the vast majority of textual features, challenge explains a significant amount of the variance in improvement. This appendix (Table C.1) lists the features with which the fitted models had significant estimated slopes, ordered by adjusted R-squared.

	Feature	β	SE	t	R^2
1	Number of verb types in their third person singular present form	0.97**	0.04	25.94	0.94
2	Standard deviation of token frequency of all words calculated with the BNC frequency list	0.91**	0.05	19.39	0.89
3	Number of cardinal number tokens.	1.38**	0.09	16.14	0.85
4	Mean token frequency of all words calculated with the BNC frequency list	0.83**	0.05	15.5	0.84
5	Number of verb tokens in their third person singular present form	0.95**	0.06	15.47	0.84
6	Number of possessive ending types.	0.87**	0.06	14.45	0.82
7	Number of Wh adjective phrases	1.04**	0.08	12.83	0.78
8	Mean type frequency of all nouns calculated with the BNC frequency list	0.93**	0.07	12.53	0.77
9	Mean token frequency of all lexical words calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.60**	0.05	10.89	0.72

	Feature	β	SE	t	R^2
10	SD of token frequency of all words calculated with the SUBTLEXus frequency list's WF measure	0.87**	0.08	10.84	0.72
11	Number of possessive ending tokens.	0.61**	0.06	10.73	0.71
12	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure	0.95**	0.09	10.46	0.7
13	SD of token frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure	0.87**	0.08	10.44	0.7
14	Number of interjection tokens.	0.68**	0.07	10.34	0.7
15	Mean token frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.56**	0.06	10.1	0.69
16	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure	0.74**	0.07	10.09	0.69
17	Number of possessive pronoun types.	1.07**	0.11	10.03	0.69
18	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure	0.96**	0.1	9.83	0.68
19	Number of plural noun tokens.	1.01**	0.1	9.8	0.68
20	number of coordinate phrases per clause	1.02**	0.1	9.79	0.68
21	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure	0.76**	0.08	9.6	0.67
22	number of coordinate phrases per T-unit	0.98**	0.1	9.41	0.66
23	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.78**	0.08	9.4	0.66
24	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure	0.96**	0.1	9.37	0.66
25	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.96**	0.1	9.27	0.65

	Feature	β	SE	t	R^2
26	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.96**	0.1	9.27	0.65
27	Mean token frequency of all lexicals calculated with the SUBTLEXus frequency list's Log10 WF measure	0.56**	0.06	9.2	0.65
28	Mean token frequency of adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.81**	0.09	9.12	0.64
29	Mean token frequency of all nouns calculated with the SUBTLEXus frequency list's WF measure	0.87**	0.09	9.11	0.64
30	Local argument overlap	1.19**	0.13	9.06	0.64
31	Mean token frequency of adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.63**	0.07	9.04	0.64
32	Mean token frequency of all words calculated with the SUBTLEXus frequency list's WF measure	0.86**	0.1	8.78	0.63
33	Mean token frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.54**	0.06	8.76	0.63
34	SD of token frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.50**	0.06	8.64	0.62
35	SD of type frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.50**	0.06	8.64	0.62
36	Standard deviation of type frequency of nouns calculated with the BNC frequency list	0.94**	0.11	8.49	0.62
37	SD of token frequency of adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.83**	0.1	8.58	0.62
38	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.83**	0.1	8.58	0.62

	Feature	β	SE	t	R^2
39	Mean token frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure	0.92**	0.11	8.54	0.61
40	Number of plural noun types.	0.73**	0.09	8.53	0.61
41	Number of verb types in their past form	0.76**	0.09	8.46	0.61
42	Standard deviation of type frequency of all adjectives calculated with the BNC frequency list	1.09**	0.13	8.35	0.6
43	Mean type frequency of all lexical words calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.64**	0.08	8.34	0.6
44	Mean token frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure	0.58**	0.07	8.31	0.6
45	Mean token frequency of all lexical words calculated with the BNC frequency list	1.05**	0.13	8.27	0.6
46	number of complex nominals per noun phrase	0.80**	0.1	8.2	0.59
47	Ratio of adjective types to lexical tokens	1.00**	0.12	8.19	0.59
48	Mean token frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.50**	0.06	8.16	0.59
49	Log-transformed type-token ratio of all verbs	0.75**	0.09	8.1	0.59
50	Number of interjection types.	0.51**	0.06	8.1	0.59
51	Mean token frequency of all lexical words calculated with the SUBTLEXus frequency list's WF measure	0.78**	0.1	8.09	0.59
52	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.89**	0.11	8.02	0.58
53	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.49**	0.06	7.96	0.58
54	Local stem overlap	0.78**	0.1	7.95	0.58
55	Mean token frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 WF measure	0.57**	0.07	7.93	0.58

	Feature	β	SE	t	R^2
56	Mean token frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.89**	0.11	7.81	0.57
57	Number of cardinal number types.	1.14**	0.15	7.73	0.57
58	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure	0.71**	0.09	7.72	0.56
59	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure	0.88**	0.12	7.63	0.56
60	Standard deviation of token frequency of adjectives calculated with the BNC frequency list	1.15**	0.15	7.63	0.56
61	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure	0.72**	0.09	7.62	0.56
62	Mean type frequency of all words calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.60**	0.08	7.59	0.56
63	Mean type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure	0.89**	0.12	7.49	0.55
64	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.77**	0.1	7.43	0.55
65	Mean type frequency of all lexicals calculated with the SUBTLEXus frequency list's WF measure	0.65**	0.09	7.42	0.54
66	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure	0.69**	0.09	7.38	0.54
67	SD of type frequency of all lexical words calculated with the SUBTLEXus frequency list's WF measure	0.67**	0.09	7.38	0.54
68	Number of verb types in their non-third person singular present form	0.79**	0.11	7.37	0.54
69	Ratio of lexical tokens to all word tokens	0.83**	0.11	7.36	0.54

	Feature	β	SE	t	R^2
70	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure	0.74**	0.1	7.27	0.53
71	Standard deviation of token frequency of all nouns calculated with the BNC frequency list	1.05**	0.15	7.16	0.53
72	Log-transformed type-token ratio of all nouns	0.73**	0.1	7.22	0.53
73	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.48**	0.07	7.21	0.53
74	Mean token frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure	0.49**	0.07	7.21	0.53
75	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.67**	0.09	7.19	0.53
76	Number of fragment clauses	0.75**	0.11	7.15	0.53
77	Mean of left embeddedness of all sentences	0.76**	0.11	7.15	0.53
78	Number of personal pronoun tokens.	0.46**	0.06	7.1	0.52
79	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's WF measure	0.90**	0.13	6.99	0.52
80	Mean token frequency of adjectives calculated with the BNC frequency list	0.87**	0.13	6.95	0.51
81	SD of left embeddedness of all sentences	0.78**	0.11	6.94	0.51
82	Mean token frequency of adverbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.87**	0.13	6.86	0.51
83	Log-transformed type-token ratio of all lexical words	0.67**	0.1	6.84	0.5
84	Number of existential there Types.	1.07**	0.16	6.83	0.5
85	Ratio of easy lexical types to all lexical types with BNC top 1000 most frequent words as easy word list	0.73**	0.11	6.83	0.5
86	Ratio of easy lexical types to all lexical types with NGSL top 1000 most frequent words as easy word list	0.73**	0.11	6.83	0.5

	Feature	β	SE	t	R^2
87	Number of fragment T-units	0.75**	0.11	6.81	0.5
88	Number of comparative adjective types.	0.94**	0.14	6.76	0.5
89	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.56**	0.08	6.76	0.5
90	Number of comparative adverb tokens.	0.65**	0.1	6.75	0.5
91	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure	0.74**	0.11	6.7	0.49
92	Mean type frequency of all lexicals calculated with the SUBTLEXus frequency list's Log10 WF measure	0.58**	0.09	6.57	0.48
93	Ratio of easy verb types to all verb types with BNC top 1000 most frequent words as easy word list	0.71**	0.11	6.46	0.48
94	Ratio of easy verb types to all verb types with NGSL top 1000 most frequent words as easy word list	0.71**	0.11	6.46	0.48
95	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.55**	0.09	6.43	0.47
96	number of complex T-units per T-unit	0.72**	0.11	6.43	0.47
97	SD of global edit distance of parse tree with part-of-speech of words	0.64**	0.1	6.42	0.47
98	Ratio of easy adverb types to all adverb types with BNC top 1000 most frequent words as easy word list	0.62**	0.1	6.41	0.47
99	Ratio of easy adverb types to all adverb types with NGSL top 1000 most frequent words as easy word list	0.62**	0.1	6.41	0.47
100	SD of global lexical overlap	0.67**	0.1	6.4	0.47
101	SD of local edit distance of parse tree with part-of-speech of words	0.70**	0.11	6.38	0.47
102	Number of superlative adverb tokens.	0.96**	0.15	6.38	0.47
103	An index that derives from the hypergeometric distribution function	0.68**	0.11	6.37	0.47

	Feature	β	SE	t	R^2
104	Guiraud's type-token ratio for all nouns	0.73**	0.11	6.36	0.47
105	Corrected type-token ratio for all nouns	0.73**	0.11	6.36	0.47
106	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.49**	0.08	6.35	0.47
107	Number of superlative adjective types.	0.99**	0.16	6.32	0.47
108	Corrected type-token ratio for all verbs	0.71**	0.11	6.29	0.46
109	Guiraud's type-token ratio for all verbs	0.71**	0.11	6.29	0.46
110	Mean type frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.59**	0.09	6.29	0.46
111	Log-transformed type-token ratio of all words	0.62**	0.1	6.28	0.46
112	Standard deviation of token frequency of all verbs calculated with the BNC frequency list	0.67**	0.11	6.25	0.46
113	Bilogarithmic type-token ratio for all verbs	0.56**	0.09	6.23	0.46
114	SD of token frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure	0.45**	0.07	6.2	0.46
115	Number of inverted declarative sentences	0.57**	0.09	6.2	0.45
116	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's WF measure	0.81**	0.13	6.18	0.45
117	Guiraud's type-token ratio for all words	0.68**	0.11	6.15	0.45
118	Corrected type-token ratio for all words	0.68**	0.11	6.15	0.45
119	Mean type frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.56**	0.09	6.15	0.45
120	Ratio of easy noun tokens to all noun tokens with BNC top 1000 most frequent words as easy word list	0.64**	0.1	6.14	0.45
121	Ratio of easy noun tokens to all noun tokens with NGSL top 1000 most frequent words as easy word list	0.64**	0.1	6.14	0.45
122	SD of global edit distance of parse tree with word tokens	0.52**	0.09	6.14	0.45

	Feature	β	SE	t	R^2
123	TTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC	1.03**	0.17	6.14	0.45
124	TTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC	1.03**	0.17	6.14	0.45
125	TTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC	1.03**	0.17	6.14	0.45
126	TTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List	1.03**	0.17	6.14	0.45
127	TTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List	1.03**	0.17	6.14	0.45
128	TTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List	1.03**	0.17	6.14	0.45
129	Number of wh-pronoun types	0.81**	0.13	6.13	0.45
130	number of clauses per sentence	0.58**	0.1	6.1	0.45
131	Number of foreign word Types.	0.89**	0.15	6.09	0.45
132	Ratio of easy noun types to all noun types with BNC top 1000 most frequent words as easy word list	0.73**	0.12	6.09	0.45
133	Ratio of easy noun types to all noun types with NGSL top 1000 most frequent words as easy word list	0.73**	0.12	6.09	0.45
134	Mean token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.44**	0.07	6.09	0.45
135	Ratio of modal tokens to all verb tokens, including modal verbs	0.88**	0.14	6.08	0.45
136	Number of comparative adjective tokens.	0.84**	0.14	6.06	0.44
137	Mean token frequency of all verbs calculated with the BNC frequency list	0.75**	0.13	6.03	0.44
138	SD of token frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.55**	0.09	5.98	0.44

	Feature	β	SE	t	R^2
139	Mean type frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure	0.56**	0.09	5.97	0.44
140	SD of token frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.59**	0.1	5.97	0.44
141	Number of wh-adverb types	0.80**	0.13	5.97	0.44
142	TTR of all words that are in the list of easy words (top 1000 most frequent) from BNC	0.77**	0.13	5.89	0.43
143	TTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC	0.77**	0.13	5.89	0.43
144	TTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC	0.77**	0.13	5.89	0.43
145	TTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.77**	0.13	5.89	0.43
146	TTR of lexicals that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.77**	0.13	5.89	0.43
147	TTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.77**	0.13	5.89	0.43
148	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 CD measure	0.84**	0.14	5.87	0.43
149	Number of superlative adverb types.	0.75**	0.13	5.84	0.43
150	Ratio of modal tokens to all word tokens	0.90**	0.15	5.83	0.43
151	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's WF measure	0.68**	0.12	5.83	0.42
152	Mean global edit distance of parse tree with lemma	0.46**	0.08	5.78	0.42
153	SD of global edit distance of parse tree with lemma	0.45**	0.08	5.77	0.42

	Feature	β	SE	t	R^2
154	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 WF measure	0.47**	0.08	5.75	0.42
155	Mean local edit distance of parse tree with lemma	0.54**	0.09	5.72	0.42
156	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure	0.83**	0.15	5.71	0.41
157	Number of foreign word tokens.	0.83**	0.15	5.7	0.41
158	Mean type frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure	0.70**	0.12	5.68	0.41
159	Corrected type-token ratio for adjectives	0.81**	0.14	5.68	0.41
160	Guiraud's type-token ratio for all adjectives	0.81**	0.14	5.68	0.41
161	Mean of local edit distance of parse tree with word tokens	0.55**	0.1	5.67	0.41
162	Number of coordinate phrases	0.75**	0.13	5.66	0.41
163	SD of token frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure	0.78**	0.14	5.66	0.41
164	Type-token ratio of all verbs	0.52**	0.09	5.63	0.41
165	Mean type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure	0.44**	0.08	5.58	0.4
166	Number of nominal clauses	0.53**	0.1	5.56	0.4
167	Number of wh-determiner tokens	0.75**	0.13	5.55	0.4
168	SD of local edit distance of parse tree with lemma	0.44**	0.08	5.54	0.4
169	Number of wh-adverb tokens	0.56**	0.1	5.53	0.4
170	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.48**	0.09	5.51	0.4
171	SD of type frequency of all lexicals calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.48**	0.09	5.51	0.4

	Feature	β	SE	t	R^2
172	Ratio of easy word types to all word types with BNC top 1000 most frequent words as easy word list	0.53**	0.1	5.48	0.4
173	Ratio of easy word types to all word types with NGSL top 1000 most frequent words as easy word list	0.53**	0.1	5.48	0.4
174	Bilogarithmic type-token ratio for all nouns	0.55**	0.1	5.47	0.39
175	CTTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC	0.74**	0.13	5.47	0.39
176	CTTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC	0.74**	0.13	5.47	0.39
177	CTTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC	0.74**	0.13	5.47	0.39
178	CTTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.74**	0.13	5.47	0.39
179	CTTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.74**	0.13	5.47	0.39
180	CTTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.74**	0.13	5.47	0.39
181	GTTR of adjectives that are in the list of easy words (top 1000 most frequent) from BNC	0.74**	0.13	5.47	0.39
182	GTTR of adverbs that are in the list of easy words (top 1000 most frequent) from BNC	0.74**	0.13	5.47	0.39
183	GTTR of verbs that are in the list of easy words (top 1000 most frequent) from BNC	0.74**	0.13	5.47	0.39
184	GTTR of adjectives that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.74**	0.13	5.47	0.39
185	GTTR of adverbs that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.74**	0.13	5.47	0.39

	Feature	β	SE	t	R^2
186	GTTR of verbs that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.74**	0.13	5.47	0.39
187	STTR of all words that are in the list of easy words (top 1000 most frequent) from BNC	0.52**	0.1	5.44	0.39
188	STTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC	0.52**	0.1	5.44	0.39
189	STTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC	0.52**	0.1	5.44	0.39
190	STTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.52**	0.1	5.44	0.39
191	STTR of lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.52**	0.1	5.44	0.39
192	STTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.52**	0.1	5.44	0.39
193	SD of local edit distance of parse tree with word tokens	0.51**	0.09	5.44	0.39
194	Number of verb types in their gerund or present participle form	0.56**	0.1	5.43	0.39
195	SD of local lexical overlap	0.60**	0.11	5.43	0.39
196	Number of determiner types	0.93**	0.17	5.41	0.39
197	Number of coordinating conjunction types.	0.70**	0.13	5.4	0.39
198	Number of Wh prepositional phrases	0.86**	0.16	5.39	0.39
199	Ratio of easy adjective types to all adjective types with BNC top 1000 most frequent words as easy word list	0.55**	0.1	5.38	0.39
200	Ratio of easy adjective types to all adjective types with NGSL top 1000 most frequent words as easy word list	0.55**	0.1	5.38	0.39
201	Mean segmented type-token ratio of all 50-word segments	0.64**	0.12	5.37	0.38

	Feature	β	SE	t	R^2
202	SD of type frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 WF measure	0.87**	0.16	5.36	0.38
203	Number of determiner tokens.	0.80**	0.15	5.36	0.38
204	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.43**	0.08	5.35	0.38
205	number of dependent clauses per T-unit	0.66**	0.12	5.35	0.38
206	Number of 'to' tokens.	0.34**	0.06	5.34	0.38
207	GTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC	0.67**	0.13	5.3	0.38
208	GTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC	0.67**	0.13	5.3	0.38
209	GTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC	0.67**	0.13	5.3	0.38
210	GTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.67**	0.13	5.3	0.38
211	GTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.67**	0.13	5.3	0.38
212	GTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.67**	0.13	5.3	0.38
213	CTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC	0.67**	0.13	5.3	0.38
214	CTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC	0.67**	0.13	5.3	0.38
215	CTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC	0.67**	0.13	5.3	0.38
216	CTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.67**	0.13	5.3	0.38
217	CTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.67**	0.13	5.3	0.38

	Feature	β	SE	t	R^2
218	CTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.67**	0.13	5.3	0.38
219	Guiraud's type-token ratio for all lexicals	0.68**	0.13	5.29	0.38
220	Corrected type-token ratio for all lexical words	0.68**	0.13	5.29	0.38
221	Type-token ratio of all adjectives	0.63**	0.12	5.28	0.38
222	SD of token frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.48**	0.09	5.28	0.38
223	Number of unknown constituents	0.50**	0.1	5.26	0.38
224	Ratio of noun types to lexical tokens	0.62**	0.12	5.25	0.37
225	Ratio of verb types to lexical tokens	0.62**	0.12	5.25	0.37
226	number of complex nominals per T-unit	0.78**	0.15	5.23	0.37
227	SD of type frequency of all nouns calculated with the SUBTLEXus frequency list's Log10 WF measure	0.92**	0.18	5.22	0.37
228	Bilogarithmic type-token ratio for all adjectives	0.68**	0.13	5.22	0.37
229	Ratio of easy lexical tokens to all lexical tokens with BNC top 1000 most frequent words as easy word list	0.54**	0.1	5.18	0.37
230	Ratio of easy lexical tokens to all lexical tokens with NGSL top 1000 most frequent words as easy word list	0.54**	0.1	5.18	0.37
231	Bilogarithmic type-token ratio for all lexicals	0.49**	0.1	5.15	0.37
232	Log-transformed type-token ratio of all adjectives	1.02**	0.2	5.15	0.37
233	Mean of global edit distance of parse tree with word tokens	0.43**	0.08	5.13	0.36
234	Number of superlative adjective tokens.	1.00**	0.2	5.11	0.36
235	SD of token frequency of adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure	1.05**	0.2	5.1	0.36
236	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's Contextual Diversity measure	1.05**	0.2	5.1	0.36

	Feature	β	SE	t	R^2
237	SD of token frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure	0.74**	0.15	5.07	0.36
238	Number of verb tokens in their gerund or present participle form	0.49**	0.1	5.04	0.36
239	Ratio of easy tokens to all word tokens with BNC top 1000 most frequent words as easy word list	0.53**	0.1	5.04	0.36
240	Ratio of all easy tokens to all tokens with NGSL top 1000 most frequent words as easy word list	0.53**	0.1	5.04	0.36
241	Mean token frequency of adverbs calculated with the BNC frequency list	0.68**	0.14	5.03	0.36
242	Number of comparative adverb types.	0.56**	0.11	5.03	0.35
243	Mean local edit distance of parse tree with part-of-speech of words	0.43**	0.09	5.03	0.35
244	Bilogarithmic type-token ratio for all words	0.46**	0.09	4.99	0.35
245	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.52**	0.11	4.93	0.35
246	Number of verb tokens in their non-third person singular present form	0.65**	0.13	4.92	0.35
247	Number of possessive pronoun tokens.	0.32**	0.06	4.92	0.35
248	number of clauses per T-unit	0.56**	0.11	4.92	0.34
249	Bilogarithmic type-token ratio for all adverbs	0.84**	0.17	4.91	0.34
250	number of verb phrases per T-unit	0.51**	0.1	4.91	0.34
251	number of complex nominals per clause	0.75**	0.15	4.91	0.34
252	Standard deviation of token frequency of all lexical words calculated with the BNC frequency list	0.56**	0.12	4.88	0.34
253	Ratio of number of adverb and adjective types to number of all lexical tokens	0.57**	0.12	4.87	0.34
254	A measure of the mean length of word strings that maintain a criterion level of lexical variation.	0.56**	0.11	4.85	0.34
255	LogTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC	0.64**	0.13	4.82	0.34

	Feature	β	SE	t	R^2
256	LogTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC	0.64**	0.13	4.82	0.34
257	LogTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC	0.64**	0.13	4.82	0.34
258	LogTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.64**	0.13	4.82	0.34
259	LogTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.64**	0.13	4.82	0.34
260	LogTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.64**	0.13	4.82	0.34
261	Mean type frequency of all verbs calculated with the BNC frequency list	0.57**	0.12	4.82	0.34
262	SD of type frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.66**	0.14	4.8	0.33
263	Type-token ratio of all lexical words	0.45**	0.1	4.78	0.33
264	number of dependent clauses per clause	0.66**	0.14	4.78	0.33
265	Type-token ratio of all nouns	0.47**	0.1	4.75	0.33
266	Number of Wh-phrases	0.46**	0.1	4.74	0.33
267	Standard deviation of token frequency of adverbs calculated with the BNC frequency list	0.79**	0.17	4.73	0.33
268	Number of verb tokens in their past form	0.44**	0.09	4.68	0.32
269	Mean length of sentence in tokens	0.53**	0.11	4.68	0.32
270	Number of determiner types.	0.57**	0.12	4.6	0.32
271	Mean global edit distance of parse tree with part-of-speech of words	0.40**	0.09	4.58	0.31
272	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.38**	0.08	4.56	0.31
273	SD of type frequency of all words calculated with the SUBTLEXus frequency list's Log10 CD measure	0.61**	0.13	4.55	0.31

	Feature	β	SE	t	R^2
274	Number of singular proper noun types.	0.45**	0.1	4.54	0.31
275	SD of type frequency of all adverbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.41**	0.09	4.51	0.31
276	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.38**	0.08	4.49	0.3
277	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's Contextual Diversity measure	0.38**	0.08	4.49	0.3
278	Mean local lexical overlap	0.56**	0.13	4.47	0.3
279	Mean token frequency of all lexical words calculated with the BNC frequency list	0.60**	0.13	4.47	0.3
280	Number of 'to' types.	0.67**	0.15	4.42	0.3
281	Ratio of easy verb tokens to all verb tokens with BNC top 1000 most frequent words as easy word list	0.55**	0.12	4.39	0.3
282	Ratio of easy verb tokens to all verb tokens with NGSL top 1000 most frequent words as easy word list	0.55**	0.12	4.39	0.3
283	Number of adjective types.	0.71**	0.16	4.35	0.29
284	Type-token ratio of all adverbs	0.76**	0.18	4.31	0.29
285	Mean type frequency of all advers calculated with the BNC frequency list	0.58**	0.13	4.31	0.29
286	Number of passive sentences	0.73**	0.17	4.29	0.29
287	Number of adjective types	0.73**	0.17	4.29	0.29
288	SD of type frequency of all words calculated with the SUBTLEXus frequency list's Log10 WF measure	0.66**	0.15	4.28	0.29
289	Mean type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.39**	0.09	4.27	0.28
290	Mean global lexical overlap	0.56**	0.13	4.24	0.28
291	Mean type frequency of all adjectives calculated with the BNC frequency list	0.65**	0.16	4.19	0.28
292	Number of punctuation mark tokens	0.21**	0.05	4.14	0.27

	Feature	β	SE	t	R^2
293	Number of adjective clauses	0.37**	0.09	4.09	0.27
294	Mean length of sentence in syllables	0.48**	0.12	4.05	0.26
295	Number of possessive wh-pronoun types	0.52**	0.13	4.03	0.26
296	Ratio of easy adverb tokens to all adverb tokens with BNC top 1000 most frequent words as easy word list	0.45**	0.11	4.01	0.26
297	Ratio of easy adverb tokens to all adverb tokens with NGSL top 1000 most frequent words as easy word list	0.45**	0.11	4.01	0.26
298	SD of token frequency of all lexical words calculated with the SUBTLEXus frequency list's Log10 WF measure	0.50**	0.13	3.99	0.26
299	Standard deviation of type frequency of all adverbs calculated with the BNC frequency list	0.50**	0.13	3.97	0.26
300	Type-token ratio of all words	0.39**	0.1	3.97	0.25
301	Guiraud's type-token ratio for all adverbs	0.62**	0.16	3.92	0.25
302	Corrected type-token ratio for adverbs	0.62**	0.16	3.92	0.25
303	Number of existential there tokens.	0.75**	0.2	3.82	0.24
304	Number of verb tokens in their past participle form	0.39**	0.1	3.8	0.24
305	Standard deviation of type frequency of verbs calculated with the BNC frequency list	0.48**	0.13	3.72	0.23
306	Number of predeterminer tokens.	0.37**	0.1	3.71	0.23
307	UberTTR of all words that are in the list of easy words (top 1000 most frequent) from BNC	0.36**	0.1	3.71	0.23
308	UberTTR of all lexical words that are in the list of easy words (top 1000 most frequent) from BNC	0.36**	0.1	3.71	0.23
309	UberTTR of nouns that are in the list of easy words (top 1000 most frequent) from BNC	0.36**	0.1	3.71	0.23
310	UberTTR of all words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.36**	0.1	3.71	0.23
311	UberTTR of lexical words that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.36**	0.1	3.71	0.23

	Feature	β	SE	t	R^2
312	UberTTR of nouns that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.36**	0.1	3.71	0.23
313	Mean length of clause in tokens	0.62**	0.17	3.61	0.22
314	Ratio of easy adjective tokens to all adjective tokens with BNC top 1000 most frequent words as easy word list	0.37**	0.1	3.6	0.22
315	Ratio of easy adjective tokens to all adjective tokens with NGSL top 1000 most frequent words as easy word list	0.37**	0.1	3.6	0.22
316	Number of passive clasuses	0.60**	0.17	3.6	0.22
317	Number of complex norminals	0.46**	0.13	3.56	0.22
318	Number of wh-pronoun tokens	0.25**	0.07	3.54	0.21
319	Number of predeterminer types.	0.35**	0.1	3.51	0.21
320	Number of verb types in their past participle form	0.35**	0.1	3.49	0.21
321	Mean token frequency of all adverbs calculated with the SUBTLEXus frequency list's WF measure	0.47**	0.14	3.42	0.2
322	Number of subordinate clauses	0.29**	0.09	3.4	0.2
323	Number of adjective phrases	0.48**	0.14	3.36	0.2
324	Number of easy adjective lemma types from the BNC list of easy words (top 1000 most frequent)	0.33**	0.1	3.36	0.2
325	Number of easy adverb lemma types from the BNC list of easy words (top 1000 most frequent)	0.33**	0.1	3.36	0.2
326	Number of easy lemma types from the BNC list of easy words (top 1000 most frequent)	0.33**	0.1	3.36	0.2
327	Number of easy lexical lemma types from the BNC list of easy words (top 1000 most frequent)	0.33**	0.1	3.36	0.2
328	Number of easy noun lemma types from the BNC list of easy words (top 1000 most frequent)	0.33**	0.1	3.36	0.2
329	Number of easy verb lemma types from the BNC list of easy words (top 1000 most frequent)	0.33**	0.1	3.36	0.2
330	Number of adjective types that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.33**	0.1	3.36	0.2

	Feature	β	SE	t	R^2
331	Number of adverb types that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.33**	0.1	3.36	0.2
332	Number of all word types that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.33**	0.1	3.36	0.2
333	Number of all lexical types that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.33**	0.1	3.36	0.2
334	Number of all noun types that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.33**	0.1	3.36	0.2
335	Number of verb types that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.33**	0.1	3.36	0.2
336	Number of sentences	0.37**	0.11	3.33	0.19
337	Number of words that are used only once	0.50**	0.15	3.32	0.19
338	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.36**	0.11	3.27	0.19
339	Number of modal word tokens.	0.56**	0.17	3.21	0.18
340	SD of token frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.36**	0.11	3.18	0.18
341	Local noun overlap	0.51**	0.16	3.17	0.18
342	Ratio of adverb types to lexical tokens	0.34**	0.11	3.15	0.18
343	Mean length of T-units in tokens	0.50**	0.16	3.15	0.18
344	Log-transformed type-token ratio of all adverbs	0.84**	0.27	3.1	0.17
345	Number of adverb types	0.44**	0.14	3.06	0.17
346	Global argument (nouns and pronouns) overlap	0.37**	0.12	3.02	0.17
347	Number of yes/no questions	0.40**	0.13	3	0.16
348	Global stem (nouns, pronouns, verbs, adjectives, and adverbs) overlap	0.42**	0.14	2.97	0.16
349	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure	0.30**	0.1	2.94	0.16

	Feature	β	SE	t	R^2
350	Number of noun types	0.42**	0.15	2.9	0.15
351	Number of verb types	0.42**	0.15	2.9	0.15
352	SD of token frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 CD measure	0.28**	0.1	2.88	0.15
353	Number of personal pronoun types.	0.30**	0.1	2.87	0.15
354	Mean type frequency of all words calculated with the SUBTLEXus frequency list's WF measure	0.46**	0.16	2.86	0.15
355	TTR of all adverbs that are not in the list of the top 2000 most frequent words of BNC	0.48**	0.17	2.79	0.14
356	Ratio of sophisticated adverb types, which are words that are not in the top 2000 most frequent words from BNC, to all adverb types	0.48**	0.17	2.79	0.14
357	TTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.48**	0.17	2.79	0.14
358	Ratio of sophisticated adverb types, which are words that are not in the top 2000 most frequent words from NGSL, to all adverb types	0.48**	0.17	2.79	0.14
359	CTTR of adverbs that are not in the list of the top 2000 most frequent words of BNC	0.46**	0.16	2.79	0.14
360	CTTR of adverbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.46**	0.16	2.79	0.14
361	GTTR of adverbs that are not in the list of the top 2000 most frequent words of BNC	0.46**	0.16	2.79	0.14
362	GTTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.46**	0.16	2.79	0.14
363	Mean type frequency of all lexical words calculated with the BNC frequency list	0.33**	0.12	2.73	0.14
364	UberTTR of all adverbs that are not in the list of the top 2000 most frequent words of BNC	0.43**	0.16	2.72	0.14

	Feature	β	SE	t	R^2
365	UberTTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.43**	0.16	2.72	0.14
366	Number of declarative clauses	0.19**	0.07	2.7	0.14
367	Number of singular proper noun tokens.	0.25*	0.09	2.69	0.14
368	Normalized type-token ratio for all verbs	0.17*	0.06	2.65	0.13
369	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 WF measure	0.43*	0.17	2.6	0.13
370	Number of easy lexical lemma tokens from the BNC list of easy words (top 1000 most frequent)	0.24*	0.09	2.55	0.12
371	Number of easy noun lemma tokens from the BNC list of easy words (top 1000 most frequent)	0.24*	0.09	2.55	0.12
372	Number of all lexical tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.24*	0.09	2.55	0.12
373	Number of noun tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.24*	0.09	2.55	0.12
374	Number of coordinating conjunction tokens.	0.33*	0.13	2.54	0.12
375	Number of Wh noun phrases	0.21*	0.08	2.53	0.12
376	Number of adverb phrases	0.36*	0.15	2.47	0.12
377	Number of sophisticated adverb tokens that are not in the list of the top 2000 most frequent words of BNC	0.36*	0.15	2.43	0.11
378	Number of adverbs tokens that are not in the list of the top 2000 most frequent words of the New General Service List	0.36*	0.15	2.43	0.11
379	Number of adverb lemmas	0.36*	0.15	2.43	0.11
380	Number of adverb tokens	0.36*	0.15	2.43	0.11
381	Number of easy lemma tokens from the BNC list of easy words (top 1000 most frequent)	0.22*	0.09	2.43	0.11
382	Number of all word tokens that are in the list of easy words (top 1000 most frequent) from the New General Service List	0.22*	0.09	2.43	0.11
383	Number of coordinate clauses	0.25*	0.1	2.41	0.11

	Feature	β	SE	t	R^2
384	Number of adverb types.	0.33*	0.14	2.38	0.11
385	Number of direct questions	0.24*	0.1	2.37	0.11
386	Number of adverb tokens.	0.33*	0.14	2.33	0.11
387	SD of type frequency of all words calculated with the SUBTLEXus frequency list's WF measure	0.38*	0.17	2.32	0.1
388	SD of type frequency of all verbs calculated with the SUBTLEXus frequency list's Log10 CD measure	0.30*	0.13	2.3	0.1
389	Number of verb phrases	0.16*	0.07	2.28	0.1
390	Number of dependent clauses	0.22*	0.1	2.22	0.1
391	Ratio of sophisticated verb types, which are words that are not in the top 2000 most frequent words from BNC, to all verb types	0.27*	0.12	2.21	0.1
392	TTR of all verbs that are not in the list of the top 2000 most frequent words of BNC	0.27*	0.12	2.21	0.1
393	TTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.27*	0.12	2.21	0.1
394	Ratio of sophisticated verb types, which are words that are not in the top 2000 most frequent words from NGSL, to all verb types	0.27*	0.12	2.21	0.1
395	SD of type frequency of all adjectives calculated with the SUBTLEXus frequency list's Log10 WF measure	0.29*	0.13	2.2	0.1
396	Normalized type-token ratio for adverbs	0.33*	0.16	2.14	0.09
397	CTTR of verbs that are not in the list of the top 2000 most frequent words of BNC	0.24*	0.12	2.07	0.09
398	GTTR of verbs that are not in the list of the top 2000 most frequent words of BNC	0.24*	0.12	2.07	0.09
399	CTTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.24*	0.12	2.07	0.09
400	GTTR of all verbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.24*	0.12	2.07	0.09

	Feature	β	SE	t	R^2
401	Number of prepositional phrases	0.32*	0.15	2.06	0.08
402	STTR of all adverbs that are not in the list of the top 2000 most frequent words of BNC	0.29*	0.14	2.03	0.08
403	STTR of all adverbs that are not in the list of the top 2000 most frequent words of the New General Service List	0.29*	0.14	2.03	0.08

β: estimated slope; SE: standard error of slope; t: t-value; $^{**}: p \leq .01$;$^{*}: p \leq .05$.

Table C.1: Detailed statistics of linear models regressing improvement on challenge with the CW corpus data