Cirsten Carlson

Elementary School L2 English Teachers' Language Performance and Children's Second Language Acquisition

UV Universitätsverlag Hildesheim

Cirsten Carlson

Elementary School L2 English Teachers' Language Performance and Children's Second Language Acquisition

Cirsten Carlson

Elementary School L2 English Teachers' Language Performance and Children's Second Language Acquisition

Universitätsverlag Hildesheim

Hildesheim
2020

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

ISO 9706
Gedruckt auf säurefreiem, alterungsbeständigem Papier
Satz und Layout: Isaias Witkowski
Umschlaggestaltung: Jan Jäger
Umschlagsabbildung: Stiftung Universität Hildesheim
Herstellung: rauer digital – druck und medien,
Markstraße 2-3, 31167 Bockenem
Printed in Germany

www.uni-hildesheim.de/bibliothek/universitaetsverlag/

ISBN 978-3-96424-025-5

Elementary School L2 English Teachers' Language Performance
and Children's Second Language Acquisition

Vom Fachbereich 3 (Sprach- und Informationswissenschaften)
der Universität Hildesheim zur Erlangung des
Grades einer Doktorin der Philosophie
(Dr. phil.)

angenommene Dissertation von
Cirsten Carlson
aus Tokio

Gutachterinnen:

Professorin Dr. Kristin Kersten
Institut für englische Sprache und Literatur, Universität Hildesheim

Professorin Dr. Nivedita Mani
Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Tag der mündlichen Prüfung: 24. September 2019

The most precious things in speech are pauses.

(Sir Ralph Richardson)[1]

1 Quoted in Crystal (2018a, p. 79).

Acknowledgements

The following work is my own, but clearly, it would have never materialized if it had not been for all the people involved during the years that it took to evolve into this final piece.

First and foremost, my heartfelt thanks go to all the teachers, students as well as their parents for allowing me to carry out the studies. It was with their fantastic cooperation that I was able to collect all the data, without which there would not have been a book.

My sincerest gratitude goes to my supervisors Kristin Kersten and Nivedita Mani. Their feedback and advice was invaluable for shaping the project from the beginning first ideas all the way to the final strokes. I am grateful for Kristin's and Nivedita's unfading support, their continuous belief and trust in my project and the time they spent. I greatly appreciate being offered the opportunity to look at my project from various academic perspectives.

Thank you to the students who gave me a hand in the data processing, of whom I would like to mention Kevin for joining me in the adventures of Praat. Another big thank you is to the colleagues who encouraged me, believed in me and helped in one way or another.

I am deeply indebted to the people who have supported me in this endeavor in so many different ways: Aga, Hannes, Kathi, Tanya, Ute, and my parents John and Monika. Their listening, their valuable thoughts, the unconditional support, and their encouragement kept me from losing sight and helped me move on.

Undoubtedly, the biggest thank you I owe the ones who were by my side every single inch of the winding rocky path: Stevie, Billie, and Neil.

Table of Contents

List of Figures

List of Tables

1 Introduction

This thesis examines elementary school L2 English teachers' language performance and children's second language acquisition. Specifically, the study focuses on public elementary school L2 English teachers and some of their students at selected schools in the state of Lower-Saxony, Germany.

Over the past years considerable effort has been made in Germany to introduce English as a foreign language into elementary school curricula. English as a foreign language has become mandatory in elementary schools in Germany starting in the third grade, as in Lower-Saxony, and in some states already in the first grade (Kultusmisterkonferenz [KMK], 2013, p. 5).[2] The number of bilingual elementary schools in Germany has increased in recent years as well. A study carried out by the association Frühe Mehrsprachigkeit an Kitas und Schulen [FMKS e.V.] counted 287 bilingual elementary schools in 2014 in Germany, which equals to 3.5 times as many as in 2003 (FMKS e.V., 2014). Included in the count were schools that offer at least one subject in a language other than German, apart from English as a subject. English-German elementary schools comprise 44% of all bilingual elementary schools (FMKS e.V., 2014). English can thus be considered the dominant language of all the languages integrated into bilingual elementary schools.

On a political level, the European Commission regards multilingualism as a fact – there are 20 official languages and about 60 indigenous languages in the European Community (European Commission [COM 596], 2005, p. 2) – as well as a goal: "The Commission's long-term objective is to increase individual multilingualism until every citizen has practical skills in at least two languages in addition to his or her mother tongue" (COM 596, 2005, p. 4).

There is virtually no disagreement over the need to learn English in Europe at some point during the school years, preferably at an early stage. However, there is no consensus on how exactly foreign language learning and teaching should take place. Primary teacher language education does not align with any single set of given objectives.

Accordingly, second language acquisition and learning have received increasing attention as research subjects. The term second language acquisition (SLA) is used "to mean the acquisition of any language(s) other than one's native language" (Larsen-Freeman & Long, 1991, p. 7). As has become customary in the field, the term SLA also includes foreign language acquisition. Albeit the learning environment and settings are different in

2 This can be considered subject to change depending on the respective current regulations and policies.

that a second language "is one being acquired in an environment in which the language is spoken natively" (Larsen-Freeman & Long, 1991, p. 6), whereas a foreign language is acquired in a context in which the target language is not spoken as a first language by the majority of the population. Foreign language acquisition is typically restricted to classroom learning (J. C. Richards & Schmidt, 2010, p. 224). Hence, the term *foreign language acquisition* will refer to classroom learning only, while *second language acquisition* includes all forms of language acquisition beyond one's first language. Thus, second language acquisition can be considered the generic term and foreign language acquisition the specific one. The present study can be considered to take place in a foreign language learning setting, as the focus is on elementary school teacher and student English in Germany, where English is taught as a foreign language at school and spoken natively only by very few. This thesis follows common practice in second language acquisition research and applies the term *second language acquisition* to any form of additional language acquisition other than the first languages.

Various academic fields investigate possibly relevant factors in language learning and acquisition: linguistics, psychology, sociology as well as their inter-disciplinary fields of psycholinguistics (e.g. Aitchison, 2011; Hatch, 1983; Traxler & Gernsbacher, 2006) and sociolinguistics (e.g. Holmes & Wilson, 2017; Hudson, 1996; Trudgill, 2000). It is fairly safe to claim that one common research finding is that language acquisition and learning is not only an innate, intrinsic process but is also highly dependent upon external factors, such as the learner's social and linguistic environment. Some research has focused on the social impact, for example Hoff (2006) and Gardner, Masgoret, and Tremblay (1999), who studied the effect of home background on language acquisition, or Sorenson Duncan and Paradis's (Sorenson Duncan & Paradis, 2018) recent study on maternal education and language acquisition. Learners' socio-economic background and its relevance to language acquisition was the focus for example in Hoff (2003). Internal factors, examined for example in Matsuda and Gobel (2004), included the relevance of anxiety in second language learning, or in Csizér and Dörnyei (2005), who examined motivation as a factor in language learning. The exact impact of those and other relevant factors, however, remain to be explored.

In all theoretical approaches to language acquisition, exposure to linguistic input is a prerequisite to language development. They differ on how much weight they attribute to linguistic input as compared to other factors affecting language development as well as how input may function in the process. Universal Grammar (UG) and usage-based approaches may be considered as two ends on the continuum of theoretical approaches to the relationship of input and language learning in such a way that UG would give input the least central place, while usage-based approaches

emphasize the driving factor of input for language acquisition. In UG, input is necessary only to trigger principles and parameters of a particular structure, after which input is no more needed (White, 2015). In usage-based approaches, on the other hand, language learning needs input to associate constructions, based on "statistical estimations" (N. Ellis & Wulff, 2015, p. 86). Input is therefore needed in large amounts, as it is believed to drive the language learning process.

Increasing demand for English instruction accompanies large numbers of teachers of English who speak English as a second language. Little is known about their language performance in general and specifically, how features of spoken language relate to the students' development of the foreign language. Explicitly or implicitly, oftentimes research is based on a conception of teachers as ideal speakers of the target language. The focus of research may then be on a variety of observations in second language learning and teaching strategies, for example on error-correction, feedback methods, or a palimpsest of methodological considerations. Publications on language teaching methodology are far too numerous to properly acknowledge here. Examples of influential publications are Celce-Murcia and McIntosh (1991), Celce-Murcia, Brinton, and Snow (2014), Harmer (2007), Nunan (1999, 2015), Richards and Lockhart (1994), and Richards and Rodgers (2014).

In second language acquisition research, however, few studies have analyzed the properties of teachers' spoken language in general and of teachers speaking the target language as a second language (L2) in particular. The thesis at hand takes teachers into consideration as second language speakers as well as agents in their students' acquisition of English as a foreign language.

The first question in developing this study was how a speakers' language performance can be analyzed in an operationalized way. Over the past two decades, a framework has been developed to describe a speaker's use of the second language in a systematic and analyzable way based on three dimensions: complexity, accuracy, and fluency (CAF). Numerous studies have applied a CAF framework to capture second language speakers' language performance (e.g. Ahmadian & Tavakoli, 2011; N. de Jong & Vercellotti, 2015; R. Ellis, 2009; Foster & Tavakoli, 2009; Housen & Kuiken, 2009; Lambert & Kormos, 2014; Larsen-Freeman, 2006; Michel, Kuiken, & Vedder, 2007; Mora & Valls-Ferrer, 2012; Muñoz, 2014; Norris & Ortega, 2009; Révész, Ekiert, & Torgersen, 2014a; Sample & Michel, 2014; Skehan, 2009; Vercellotti, 2015; Wolfe-Quintero, Inagaki, & Kim, 1998; Yuan & Ellis, 2003). A variety of measures used in the framework underpinning each of the CAF dimensions have become frequent means to make statements about an L2 speaker's language.

Yet, findings in the field of complexity, accuracy, and fluency are not conclusive, and respective findings support different theoretical considerations about language production. Specifically, it has been a matter of debate whether L2 speakers can perform equally well on all three CAF dimensions at the same time, or if the dimensions come at the expense of one another. Skehan (1998, 2009) and Robinson (2003, 2011) have formulated according hypotheses on language production, namely the Limited Capacity Hypothesis, also referred to as Trade-Off Hypothesis, and the Cognition Hypothesis. The hypotheses predict L2 speakers to perform unbalanced on the CAF dimensions. The Limited Capacity Hypothesis expects the dimensions to compete due to limited attentional and working memory capacity of second language speakers. The Cognition Hypothesis also predicts competition in the performance on the CAF dimensions. Research is not clearly decided on whether the dimensions inevitably trade off and if so, which ones come at the expense of one another.

By applying the CAF framework in the present thesis, several objectives were followed. First, the framework served as a means to measure language performance of the teachers as part of their over-all L2 proficiency. Second, the results added to an understanding of how complexity, accuracy, and fluency may relate to one another.

The second main question in approaching the thesis topic was how the teachers' linguistic L2 performance can be related to their students' L2 development empirically. Studies on early second language acquisition have frequently based the assessment of children's L2 development on receptive skills in two language areas, vocabulary and grammar, by using standardized tests (e.g. Buyl & Housen, 2015; Couve de Murville, Kersten, Maier, Ponto, & Weitz, 2016; Hopp, Kieseier, Vogelbacher, & Thoma, 2018; Horváth & Nikolov, 2007; Jaekel, Schurig, & Florian, 2017; Maier, Neubauer, Ponto, Couve de Murville, & Kersten, 2016; Rohde, 2010; Schelletter & Ramsey, 2010; Steinlen, Håkansson, Housen, & Schelletter, 2010; Steinlen & Piske, 2016; Steinlen & Rogotzki, 2008; Unsworth, Persson, Prins, & Bot, 2015). Such test results can be considered an indication of the children's L2 development. A statistical model could then relate the teachers' language performance to the children's test results.

The study is a mixed methods approach with the following design: The first strand of the study – Study 1 – is an analysis of eleven elementary school English teachers' performance in English. The study focuses on how linguistic performance as part of overall proficiency can be measured. Complexity, accuracy, and fluency (CAF) have increasingly been considered core dimensions of linguistic performance and, as a framework, been applied to describe second language development. In order to capture language performance in an operationalized manner, batteries of quantitative measures have been developed alongside for each of the

three dimensions. However, inconclusive study findings suggest to further develop approaches to operationalizing CAF. In the present study, the CAF framework is used to give an account of the spoken language features of the teachers' English using semi-guided interviews. The main research question guiding Study 1 is as follows: (RQ1) How do the L2 English teachers perform considering complexity, accuracy, and fluency?

The study design focused on the teachers' language performance as part of their L2 proficiency and did therefore not include classroom observation. However, as teachers may modify their language in the classroom, a questionnaire substudy was included to add information on the teachers' language use in the classroom, examining the following research question: (RQ2) How do the teachers rate their L2 English language proficiency and the modification of their language use in the classroom?

The second strand of the empirical study – Study 2 – looks at how the receptive English of 132 elementary school students of a subset of four of the eleven teachers develops. Receptive vocabulary and grammar are chosen as indicators of the children's early stages of foreign language development, when productive skills are little developed. The study is based on two standardized tests: the British Picture Vocabulary Scale 3 (BPVS3) (Dunn, Dunn, Styles, & Sewell, 2009) for receptive vocabulary and the ELIAS Grammar Test II (Kersten, Piske, et al., n.d.) for receptive grammar. Study 2 focuses on the following research questions: (RQ3) How do the students' receptive English grammar and vocabulary develop over their fourth year of elementary school? (RQ4) How do the student groups differ per teacher in their receptive English vocabulary or grammar attainment and development?

The third strand of the empirical chapter consists of Study 3, a synthesis of both study strands. It first aims to contribute to an understanding of the relationships between complexity, accuracy, and fluency and to develop a procedure that can operationalize the performance dimensions for further analysis. Second, the synthesis study investigates the teachers' linguistic performance as a possible factor in second language acquisition. It explores whether a connection can be detected between the teachers' measured oral production and the students' outcomes on the vocabulary and grammar tests.

Therefore, the synthesis Study 3 is divided into two parts. Part I of the synthesis study applies a novel procedure to capture CAF in operationalized scores for each dimension. This part examines the following research questions: (RQ5) How can the CAF dimensions be transformed into a scale that can be used for further analyses? (RQ6) How do complexity, accuracy, and fluency in the teacher's L2 performance relate to one another?

Finally, Part II of the synthesis Study 3 joins four of the teachers' CAF performance and their students' L2 development, examining the following research questions: (RQ7) How does the teachers' L2 English performance,

as measured in complexity, accuracy, and fluency, relate to their students' L2 receptive vocabulary and grammar development? (RQ8) If there is a relationship between teachers' L2 performance and children's foreign language acquisition, is there an additional effect by the classroom L2 use as rated by the teachers?

The outline of this thesis is as follows: Chapter 2 comprises the theoretical background needed to underpin the motivation for the studies. The dimensions of language performance and the operationalization of complexity, accuracy, and fluency in language production are reviewed and discussed first. Because there still is need to clarify the relationships between those three dimensions, the state of the art regarding the interrelationships is reported as well.

Following is a section on linguistic input, in which the term in use is presented, as well as the state of the art concerning whether and which features in the input may affect second language acquisition. A brief examination of first language acquisition research regarding linguistic input factors supplements insights on input effects, as do language instructional particularities such as teacher-talk and teaching strategies.

Chapter 3 presents the empirical study. The teachers' linguistic performance in interviews is examined in Study 1. The data elicitation, coding of the transcripts, and the measures indexing complexity, accuracy, and fluency as performed in the interviews are explained and analyzed. A substudy introduces a questionnaire, asking those four teachers, whose classes took part in the testing of grammar and vocabulary, for self-ratings on their target language. Study 2 follows, studying students' receptive grammar and vocabulary in nine classes taught by four of the interviewed teachers at four different schools. The overall group scores are calculated as well as the individual student scores and categorized by the teachers who taught them. The synthesis Study 3 is subdivided into two parts. Part I analyzes the CAF measures applied in teacher interview Study 1, the measures' mutual relationships, and the contribution of the measures to each respective CAF dimension in a Principal Component Analysis (PCA). CAF factor scores were calculated based on the PCA results in order to obtain composite scores of each complexity, accuracy, and fluency dimension in the language production of each of the eleven interviewed teachers. In Part II of the synthesis Study 3, the scores of the subset of four teachers were used in the final analysis of relating the teachers' linguistic performance to their students' L2 development. Regression analyses calculated whether any of the CAF dimensions as measured in the teachers' performances predicted the children's receptive grammar or vocabulary development.

The conclusion summarizes the main findings, states the limitations of the studies, and discusses the results' implications for second language teaching and future research.

2 Theoretical Background and State of the Art

In order to investigate language performance and understand possible relationships between teachers' L2 language and their students' second language acquisition, several theoretical considerations need to be taken into account first. The following sections look at the terms and concepts in use – linguistic performance and how it can be operationalized, linguistic input and where it is positioned in first and language acquisition research, how and which teaching strategies are considered beneficial with respect to the acquisition of a second language, and finally, how children's L2 acquisition is assessed in two areas representing language acquisition – receptive grammar and vocabulary.

2.1 Dimensions of Language Performance

The teacher participants of the current study were L2 English speakers living in a dominantly German-speaking environment. They were therefore second language speakers as well as teachers. As there are no formal target language requirements to become an English teacher at elementary level in Germany, there is also no proficiency baseline that can be expected. There are several ways to become an elementary school English teacher in Germany: one is to earn a teaching degree in English studies in addition to other majors. By rule, graduates of a teaching degree in English are expected to have a European Framework C1 level of English. However, there is no mandatory language testing for future teachers in place. Graduates may differ considerably in their proficiencies of English. A second possibility to become a primary level English teacher is to undergo further formal training in order to qualify for a teaching career. Those participants are also expected to be at a C1 level of English (Kultusmisterkonferenz KMK, 2013, p. 8). Yet another possibility is to teach English on demand despite lacking a degree in English. The responsibility to assign the teaching schedules and subjects lies with the schools. Because teachers at the elementary level may teach every subject, the number of English teachers without a degree in English is considered far higher than the number of those holding a degree in English: In an interview, Piske (2011) estimates that about 75% of the elementary school English teachers in Germany do not hold a degree in English. Consequently, the elementary school English teachers in Germany can vary greatly in their English speaking skills, as a particular level of English is not a given. As a result, next to no claims can be made about

the elementary teachers' target language proficiency. For the current study, it is necessary to take into consideration what exactly can be measured in language performance and how it can be operationalized.

Any approach that aims to capture any form of language proficiency will be concerned with the question how proficiency can be determined: "What makes a second or foreign language (L2) user, or a native speaker for that matter, a more or less proficient language user?" (Housen & Kuiken, 2009, p. 1). Housen, Kuiken, and Vedder (2012a) summarize that the principal elements of L2 proficiency "can be fruitfully captured by the notions of complexity, fluency and accuracy" (p. 1).

The complexity, accuracy, and fluency (CAF) framework was developed in order to capture what is believed to be a multidimensional process of language performance and to describe a speaker's use of the second language in a systematic and analyzable way. The notions of complexity, accuracy, and fluency have been used in research for roughly the past twenty years to describe language learners' production (e.g. Ahmadian & Tavakoli, 2011; R. Ellis, 2009; R. Ellis & Barkhuizen, 2005; R. Ellis & Yuan, 2004; Ferrari, 2012; Foster & Tavakoli, 2009; Sample & Michel, 2014; Skehan, 1998; Vercellotti, 2015; Wolfe-Quintero, Inagaki, & Kim, 1998; Yuan & Ellis, 2003).

Complexity, accuracy, and fluency can be considered descriptors, or dimensions, of language production. Even though performance and proficiency are sometimes used as synonyms (e.g. Housen, Kuiken, & Vedder, 2012b), CAF are in fact indicators, or components, of language performance, which in turn feed into proficiency. Linguistic performance in terms of CAF measures language production in those three domains. Language proficiency on the other hand includes a much broader sense of language use: "Proficiency is the ability to use language in real world situations in a spontaneous interaction and non-rehearsed context and in a manner acceptable and appropriate to native speakers of the language" (*ACTFL Performance discriptors for language learners*, 2012, p. 4). Such a notion of proficiency includes aspects of speaking, listening, writing, and reading skills along with pragmatic and discourse behavior in the second language. The relationship between proficiency and performance is described by the American Council of the Teaching of Foreign Languages (ACTFL) as follows:

> Demonstration of performance within a specific range may provide some indication of how the language user might perform on a proficiency assessment and indeed might point toward a proficiency level, but performance is not the same as proficiency. (*ACTFL Performance discriptors for language learners*, 2012, p. 4)

According to a number of studies, performance measures of complexity, accuracy, and fluency predict language proficiency as tested or perceived by raters (Iwashita, Brown, McNamara, & O'Hagan, 2008; Révész, Ekiert, & Torgersen, 2014b; Seedhouse, Harris, Naeb, & Üstünel, 2014). Those results support the idea of complexity, accuracy, and fluency as being an integral part of an overall language proficiency, albeit not the same as proficiency. The field of language testing, however, can be considered one of its own, as it is of particular interest to language testing systems and their agencies.

Throughout the current paper, *performance* relates to spoken language production as indicated by its complexity, accuracy, and fluency. When the term *language proficiency* is used hereafter, it is either in terms of an overall language proficiency that may include various other unspecified aspects, or when the corresponding sources have used it.[3]

As components of language production, complexity, accuracy, and fluency have each been in use to describe language use in various distinctive ways since the 1990s, when Skehan (1998) brought the three dimensions together. Skehan (2009) refers back to Crookes (1989) to arguably have been one of the first investigating similar dimensions to what has now become known as the CAF framework. Those three dimensions have become a working triad to capture language performance based on the following broad working definitions: Complexity refers to "the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2" (Housen et al., 2012a, p. 2). Accuracy is considered "the ability to produce target-like and error-free language" (Housen et al., 2012a, p. 2). Fluency then is "the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation" (Housen et al., 2012a, p. 2). All three definitions have been subject to debate in particular with regard to notions such as 'sophisticated', 'norm', 'native-like' or 'error-free' (e.g. Pallotti 2015) – a theoretical discussion which particularly determines the choice of measures of each dimension.

Despite remaining theoretical as well as practical inconsistencies, the CAF notions have gained momentum more recently not only as descriptors of language performance, but also as a theoretical framework of language production (e.g. Pallotti, 2009; Norris & Ortega, 2009; Housen et al., 2012a; Pallotti, 2015; Vercellotti, 2015). *Figure 1* illustrates the three dimensions of performance.

3 For a discussion of other aspects of language proficiency such as sociolinguistic or discourse skills, see Harley, Allen, Cummins and Swain (1990) and Leclerq, Edmonds, and Hilton (2014).

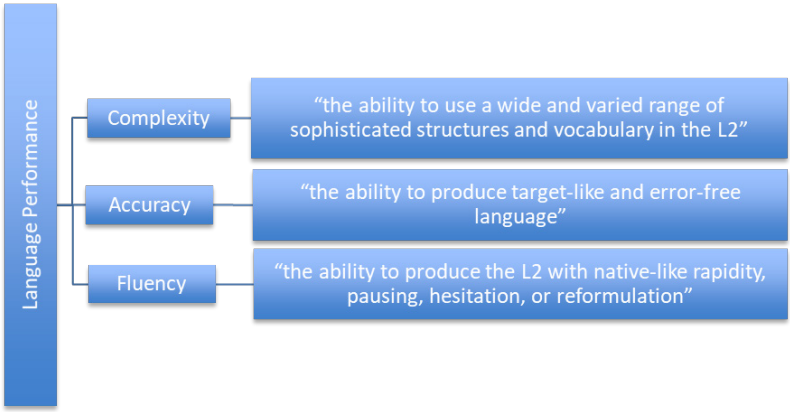

Figure 1 CAF triad
Quotes: Housen et al. (2012a, p. 2)

In *Figure 1* complexity, accuracy, and fluency are shown at the same level. Based on Skehan's (1998) considerations on the relationship between the three dimensions, R. Ellis and Barkhuizen (2005) place a form and meaning level between language performance and the three dimensions, which subscribes fluency to the meaning level, and complexity and accuracy to the form level in language performance.

Analyzing linguistic data according to the CAF framework is based on a number of set sequences and indices. The aim of such a detailed linguistic analysis is to capture the complexity, accuracy, and fluency of language production validly, reliably, and feasibly.

As each dimension of complexity, accuracy, and fluency is a concept in itself and is applied in the empirical study at hand, all three need to be discussed in more detail in the following sections. Each dimension is also based on a number of measures, which are discussed in section 2.1.4 and form the base for the respective data analysis of the current empirical study.

2.1.1 Complexity

Of the three dimensions complexity, accuracy, and fluency, "[c]omplexity is the most complex, ambiguous and least understood," Housen and Kuiken (2009, p. 4) state. The definition of complexity as "the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2" (Housen et al., 2012a, p. 2) intends to be general rather than specific. Pallotti (2009), too, calls complexity the "most problematic" (p. 5) dimension in the CAF triad because of its "polysemous nature" (p. 5). Part of the problem

is that the term *complexity* has been used in different ways entailing task as well as language performance properties, indicating objective versus subjective difficulty.

Pallotti (2009) summarizes that there are at least three ideas of the term *complex* found in research. The first notion of complexity is plainly structural. It relates to the respective inherent linguistic structure in terms of the number of parts or components, for example one transformational rule versus multiple transformational rules. Linguistic variety, this is the number of alternative forms used in language, also underlies this structural idea of complexity and frequently defines lexical diversity as well (Pallotti, 2009, p. 6f.). The second idea of complexity equates with difficult, cognitively demanding or "challenging language" a speaker can attend to using – an understanding found in Skehan (2009, p. 511), for example. The third sense of complexity is related to the acquisitional stages, equaling *complex* and "more advanced", for example in R. Ellis (2009, p. 2). All three ideas can overlap, but as Pallotti (2009) sums up, "this correspondence has to be demonstrated, not assumed, and there is no guarantee that it holds universally" (p. 7).

The concept of complexity can further be divided into linguistic complexity and cognitive complexity, as Housen et al. (2012a) explain: "Linguistic complexity is an objective given, independent from the learner, which refers to the intrinsic formal or semantic-functional properties of L2 elements (e.g. forms, meanings, and form-meaning mappings) or to properties of (sub-) systems of L2 elements" (p. 4). Cognitive complexity, on the other hand, "is a relative and subjective notion" and refers to the "relative difficulty with which language elements are processed" (Housen et al., 2012a, p. 4). However, even if the use of complexity refrains from including a cognitive notion and is restricted to "performance description, it still retains multiple meanings, because it can be applied to different aspects of language and communication" (Pallotti, 2009, p. 6).

Another uncertainty when speaking about complexity pertains to whether the notion of complexity has to include a developmental idea, which relates to when or at what developmental stage specific structures may or may not be acquired. Pallotti (2009) argues to keep language development separate from CAF as development involves a process. Instead, he suggests that CAF "refer to the properties of language performance as a product" (Pallotti, 2009, p. 8). In the same vein, Pallotti (2015) later argues to follow a simple view of complexity:

> In order to avoid this polysemy, this article advocates a simple view of complexity, treating it as a purely descriptive category, limiting its use to structural complexity and excluding from its definition any theoretical

> assumption about when, how and why it increases or remains constant. (Pallotti, 2015, p. 119)

Since the study at hand is interested in giving an account of the linguistic performance of the participants, it is such a "simple view" that forms the theoretical base of the complexity dimension. The study aims to capture the study subjects' linguistic performance and its properties as a product and does not measure the development of particular linguistic features.

Initially not being part of the CAF triad, a lexical dimension was later called to be included. Skehan (2009), for example, strongly criticized the absence of a lexical dimension in task research. In his view it was "a serious omission" (p. 514), and he argued as follows:

> The lexis-syntax connection is vital in performance models […] such as Levelt's, and lexis represents a form of complexity that has to be assessed in second language speech performance if any sort of complete picture is to be achieved. (Skehan, 2009, p. 514)[4]

Skehan therefore argued to supplement the three CAF dimensions by a lexical measure. Lexis, when considered, has then been attributed to the notion of complexity in the CAF triad (Skehan, 2009). Lexis is considered a form of complexity and is therefore referred to as lexical complexity (R. Ellis & Barkhuizen, 2005). According to Pallotti's advocated "simple view" of complexity, "complexity can be operationalized essentially in terms of diversity. A text with a wide variety of lexemes will be said to be more complex than one where the same few words are repeated over and over" (Pallotti, 2015, p. 125).

Likewise, a text showing more subordination will be considered more syntactically complex. However, lexical complexity as used and defined does not in any way inform about the complexity of words, for example in terms of their frequency or pronunciation difficulty. It rather refers to the variety of words used. Thus, lexical complexity indicates lexical variety, or more precisely lexical diversity. The dimension of complexity is typically divided into syntactic complexity and lexical diversity (e.g. N. de Jong & Vercellotti, 2015; R. Ellis, 2009; R. Ellis & Yuan, 2004; Ferrari, 2012; Foster & Tavakoli, 2009; Inoue, 2010; Lintunen & Mäkilä, 2015; Muñoz, 2014; Révész, Ekiert, & Torgersen, 2014a; Sample & Michel, 2014; Vercellotti, 2012, 2015; Verspoor, Lowie, Chan, & Vahtrick, 2017; Yuan & Ellis, 2003).

4 Skehan refers to Levelt's model of language production (Levelt, 1989).

2.1.2 Accuracy

At first sight, accuracy may seem to be the most easily definable dimension in the CAF triad. Accuracy "refers to the extent to which an L2 learner's performance […] deviates from a norm (i.e. usually the native speaker)" (Housen et al. 2012a, p. 4). According to this working definition, any deviation is considered an error.

At second sight, however, severe difficulties become prominent when dealing with language as a dynamic system and with its speakers as multifaceted and developing individuals. The first problem not entirely solvable is the concept of a norm as it depends on the part of world and the variety considered acceptable there. Native speakers of different varieties of English may consider different language features acceptable according to the norm of that variety. This is the case in certain lexical items, for example *pavement* and *sidewalk* in the UK and North America respectively, as well as in grammaticality judgments, for example the past tense forms *fitted* versus *fit*, the former being the standard past tense in the UK and the latter the standard form in North America ("fit," n.d.).[5] Likewise, two speakers of an English variety from the same country, but with different socio-economic, ethnic, or linguistic backgrounds, might follow differing standards of the acceptability of a given form as well. In addition, the native speaker norm concept has become a matter of debate in light of the world's majority of English speakers being L2 speakers. Cook's (2006) opinion illustrates how far acceptability of non-native speech may range when L2 speakers are considered different instead of deficient: "L2 users have the right to speak English as L2 users rather than as imitation native speakers" (p. 50). [6]

Of all languages, particularly English as a lingua franca in many international contexts is subject to an enormous amount of change. For example, forming a plural form in nouns which are non-count nouns in any standard native English variety, such as *informations* or *advices,* has become a commonplace practice in international contexts of English as a lingua franca (Seidlhofer interviewed in Skapinker, 2007). Several researchers have suggested that because English is the lingua franca in many parts of the world and most international communication, English as a global language requires to be dealt with as a natural language rather than a faulty copy of any native English variety (e.g. Crystal, 2003; Jenkins, 2015; Seidlhofer, 2013). When accuracy is the object of assessment, however, "the purpose of accuracy measures is precisely the comparison with target-like use", as Wolfe-Quintero et al. (1998, p. 33) point out.

5 For more on varieties of English see Hughes, Trudgill, and Watt (2013), Mesthrie and Bhatt (2008), and Trudgill and Hannah (2013).

6 In this dissertation, the term *non-native* is used interchangeably with the term *L2.*

The second difficulty in determining accuracy is to decide what is considered an error or deviation and whether errors can be more, or less accurate, depending on the respective norm. For this reason, Housen et al. (2012a, p. 4) argue that accuracy be understood as appropriateness and acceptability.

Pallotti (2009), on the other hand, argues against differentiating among error types when the concept of measurable accuracy is concerned. He reasons that

> a 100-word production with ten errors not compromising communication is not more 'accurate' than a text of the same length with ten errors hindering comprehension, but just more 'understandable' or 'communicatively effective'. (p. 5)

Pallotti (2009) therefore proposes considering adequacy a separate dimension from accuracy and illustrates the difficulty of defining accuracy in terms of adequacy in the following two statements: *colorless green ideas sleep furiously*, a sentence originally introduced by Chomsky (1957), which is grammatically accurate, but communicatively inadequate, and *me no likes go dance* (Pallotti, 2009, p. 5). The latter is likely to be perfectly intelligible, but deviates from various English grammar norms. So an utterance can be communicatively acceptable even though it is against the rules of a defined standard norm. Likewise, a sentence can be unintelligible even though it is grammatically correct.

Along similar lines, Pallotti (2009) agues not to mingle accuracy and development of a language by weighing errors according to their developmental sequence: "A 100-word text with ten errors on subjunctives and conditionals is not 'more accurate' than one with ten errors on articles and pronouns, but simply 'more developed' or 'advanced'" (p. 5). He concludes that a distinction should be made between accuracy and development. Both ought to be treated separately, requiring different measures when assessed.

Taking up "the old chestnut" (Foster & Wigglesworth, 2016, p. 105) of the phrase *colorless green ideas sleep furiously,* Foster and Wigglesworth (2016) argue from a practical perspective pointing out that

> because research participants have ordinary instincts for communication, an analysis of spoken or written data should be able to start from the assumption that the speaker or writer is not trying to be deliberately obscure, irrelevant, untruthful, or ambiguous. (p. 105)[7]

7 They refer to Grice's cooperative principles (Grice, 1975).

Foster and Wigglesworth (2016) therefore call for a differentiated system of weighing errors. They introduce a weighted clause ratio (WCR), which evaluates errors on a four-level scale according to their impact on communication, ranging from entirely accurate clauses to clauses seriously impeding meaning. Rater subjectivity in deciding which errors impede meaning and which do not, however, is built in, particularly if the raters speak the language of the participants.

Despite disagreements in the field on defining accuracy, the commonality is that accuracy is most often defined based on morpho-syntactical and lexical levels. Pronunciation as a feature in language production, however, is usually not considered because considering pronunciation as part of the accuracy dimension would pose another problem, as the varieties of English differ most notably in their accents. Deciding on the accuracy of pronunciation is "especially problematic given the difficulty in determining what constitutes the appropriate target accent to use as a baseline for examining the learner's pronunciation" (Ellis & Barkhuizen, 2005, p. 151f.). To my knowledge, no study investigating CAF dimensions has incorporated pronunciation into the accuracy dimension, or any other dimension for that matter. For the study at hand, accuracy is based on error-freeness in the participants' speech, in accordance with Pallotti's (2009) understanding of structural accuracy.

2.1.3 Fluency

Of the three dimensions, fluency can be considered the most manifold one in terms of the number of definitions and measures that may apply. Fluency is one of the prominent factors that account for a person's proficiency in a language as perceived by other speakers of the language. Language testing and assessing will usually include fluency of some sort as a determinant in rating spoken performance. The can-do-statements of the CEFR include comments on the speaker's fluency: for example, on the global scale of the CEFR a C2 speaker "[c]an express him/herself spontaneously, very fluently and precisely" (Council of Europe, 2001, p. 24).

In order to disentangle the understandings of fluency, Lennon (1990) distinguishes two basic notions of fluency – a broad sense and a narrow sense. Fluency in a broad sense indicates a speaker's general oral proficiency in a language. In this sense, fluency is a cover term that might include all areas of language production. According to the broad sense, being a fluent speaker equals being a highly proficient speaker of a language, using a large vocabulary, accurate complex grammar, and possibly some sort of native-like pronunciation. The narrow sense of fluency refers to one component of oral performance, which captures the smoothness and apparent ease

of language production. It is a component that is frequently at the core of language assessment such as in "'fluent but grammatically inaccurate'" (Lennon, 1990, p. 390). Underpinning this narrow sense of fluency is a notion of the rapidity in language production manifested in pausing behavior, hesitations, repetitions, and self-repairs.

The broad and the narrow sense of fluency, however, interplay: "The implication is that it is fluent delivery in performance that is probably the overriding determiner of perceived oral proficiency" (Lennon, 1990, p. 391). Accuracy and other features would then be of lesser importance than perceived spoken fluency.[8]

Segalowitz (2010, 2016) distinguishes three aspects of fluency – utterance fluency, cognitive fluency, and perceived fluency. Utterance and cognitive fluency are similar to what was earlier described by Lennon (1990) as the narrow sense of fluency. Utterance fluency "refers to the fluidity of the observable speech as characterized by measurable temporal features such as syllable rate, duration and rate of hesitations, filled and silent pauses, and including what Skehan (2003) has identified as breakdown fluency and repair fluency" (Segalowitz, 2016, p. 81).

Utterance and cognitive fluency relate to the automaticity with which language is produced. In his model, Segalowitz (2016) explains cognitive fluency features to refer "to the fluid operation (speed, efficiency) of the cognitive processes responsible for performing L2 speech acts" (p. 82).

Studies in the field of linguistic language performance and second language fluency as a dimension in the CAF triad are concerned with the narrow sense of utterance fluency (Ahmadian & Tavakoli, 2011; Bosker, Pinget, Quené, Sanders, & de Jong, 2012; N. H. de Jong, 2016; N. H. de Jong, Groenhout, Schoonen, & Hulstijn, 2015; N. H. de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013; Ferrari, 2012; Inoue, 2010; Lambert & Kormos, 2014; Pinget, Bosker, Quené, & de Jong, 2014; Révész et al., 2014a; Sample & Michel, 2014; Vercellotti, 2012).

Three sub-dimensions have been identified to include utterance fluency variables: (a) speed fluency, i.e. the rate and density of language production, (b) breakdown fluency, i.e. the number, length and location of pauses, and (c) repair fluency, i.e. self-repairs, repetitions, reformulations (Skehan, 2003). As a dimension in the CAF triad, fluency is understood in the narrow sense and can be summarized as follows: "Fluency is primarily related to learners' control over their linguistic L2 knowledge, as reflected in the speed and ease with which they access relevant L2 information" (Housen & Kuiken, 2009, p. 3). Since the present study focuses on measuring fluency features as part

8 It is noteworthy that the picture is different for written language, where the opposite may be the case: accuracy rather than fluency or any other aspect would predominantly determine the assessment (Lennon, 1990, p. 391).

of the speakers' performance on the three CAF dimensions, it is primarily interested in the narrow sense. Therefore, fluency refers to the narrow sense of utterance fluency as determined by features of speed, breakdown, and repair fluency.

2.1.4 Operationalizing Language Performance

The following sections look at how language performance can be measured in an operationalized way, discusses the measures, and sets the background for the empirical study on language performance in section 3.2. The basic problem of how language data can be sequenced is reviewed and discussed first, followed by a report and critical evaluation of the measures applied to quantify complexity, accuracy, and fluency.

2.1.4.1 Sequencing Spoken Data

Contrary to written data, oral data does not indicate unit boundaries by punctuation marks such as periods. In addition, speech often presents incomplete sentence structures, which makes identifying unit boundaries enormously challenging. Yet al. calculations of indices are based on sequencing the linguistic material into quantifiable units. Approaches to sequencing spoken language can be found early in research on first language acquisition. As babies' and toddlers' emerging language production is by nature fragmentary, other operationalizable units than sentences were needed.

Mean Length of Utterance (MLU), introduced by Brown (1973) to segment language of beginning first language speakers, was considered not well applicable to adult speakers (see Crookes, 1990). Intonation units cannot be considered reliable either, as Crookes (1990) reports, for example when foreign language speaker data is involved. English L2 speakers might follow different intonation patterns than native speakers of the language. A "neutral intonation pattern", which Crookes (1990, p. 190) refers to as a characteristic of the English variety Edinburgh Scottish English, can also be found in some L2 English speakers. Many standard English varieties, however, do not follow a neutral intonation pattern.

A unit frequently found in language data analysis is the T-unit, or the minimal terminal unit, which traces back to Hunt (1965). It was predominantly developed for the analysis of written language and is based on sentence structure, but has been used for oral data as well. A T-unit is an independent clause and any of its associated dependent clauses. It is therefore syntactically based. In the following examples, only (a) and (b) are T-units, whereas (c) is not:

(a) John woke up
(b) John woke up, although he was tired
(c) although he was tired (Gass, 2013, p. 65)

The T-unit has been used in operationalizing written performance of second language learners (e.g. Hunt, 1965; Larsen-Freeman, 2006, 2009; Wolfe-Quintero et al., 1998; Yuan & Ellis, 2003). Even though T-units have not only been in use for written language but also for spoken data (e.g. Bygate, 2001; R. Ellis & Yuan, 2004; Kawauchi, 2005; Larsen-Freeman, 2006; Lennon, 1990), T-units are considered most reliable with written data (Gass, 2013, p. 65). Because of their reliance on sentences, T-units were not rendered reliable for spoken learner data as oral data frequently lack complete sentence structures and include short utterances or incomplete fragments. As a result of applying T-units to spoken language, a large amount of the data in spoken language that cannot be segmented according to a syntactical unit alone may then be excluded from analysis.

The communication unit (C-unit) expands the T-unit to accommodate for spoken language features, such as fragmentary speech, and was first introduced by Loban (1966). The communication unit was designed specifically to cope with fragmentary oral data and was later used for example in Foster and Skehan (1996). C-units include in the analysis short utterances such as *yes, no, uh-huh* and short utterances without verbs. That way, elliptical utterances could be included in the analyses. However, elliptical speech taking place within the same speaker's utterance could not be accounted for in C-units. The C-unit was considered more valuable to the analysis of spoken language than the T-unit, but it lacked a more detailed definition to account for all spoken data (Skehan, 2003).

In an effort to unify segmentation units for language data analysis, Foster et al. (2000) first examined how spoken data were most often analyzed. They found a severe lack of unifying terms and definitions when it came to spoken language analysis. Of the 87 studies Foster et al. (2000) searched for unit definitions, 44 did not define the unit their analyses were based on. Foster et al. (2000) further examine the shortcomings of the T-unit and the C-unit as measures to analyze spoken data and eventually incorporated earlier segmentation units into a new unit, the analysis-of-speech unit (AS-unit).

The AS-unit is, like the T-unit, primarily syntactically based, but it is complemented by independent sub-clausal units found in spoken language (Foster et al., 2000, p. 366). The AS-unit intends to include levels of details that are needed for the segmentation of spoken language. The clause was defined as a syntactic subunit "allowing a more finely grained analysis of oral data by basing it on something that is likely to be shorter, and thus less prone to error, than a complete unit" (Foster & Wigglesworth, 2016, p. 102).

Foster et al. (2000) define AS-unit as follows: "An AS-unit is a single speaker's utterance consisting of an *independent clause, or sub-clausal unit,* together with any *subordinate clause(s)* associated with either" (p. 365). They further define an independent clause to include a finite verb and an independent sub-clausal unit to consist of

> *either* one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation:
> A: /how long you stay here /
> B: /three months/
> *or* a minor utterance, which will be defined as one of the class of 'Irregular sentences' or 'Nonsentences'. (Foster et al., 2000, p. 366)

Examples of nonsentences are the following utterances:

> /Oh poor woman/
> /Thank you very much/
> /Yes/ (Foster et al., 2000, p. 366)

According to Foster et al.'s measures, the following utterance is counted as two clauses and one AS-unit:

> /it is my hope :: to study crop protection/ (p. 366)

As opposed to basing measures on a random word count, "segmentation of units at syntactic boundaries does have a claim to psycholinguistic reality […] rendering units such as a clause, a T-unit, and an AS-unit stronger candidates for the basis of an analysis" (Foster & Wigglesworth, 2016, p. 102).

By now, a shift towards using AS-units for spoken language analysis in the field of second language research has emerged so that Foster et al.'s (2000) call in their article title to have "a unit for all reasons" seems to have been heard. The AS-unit has become the unit that is considered the most reliable one and the unit most frequently used in CAF analyses (e.g. N. H. de Jong, 2016; N. H. de Jong & Bosker, 2013; Ferrari, 2012; Inoue, 2010; Révész, Ekiert, & Torgersen, 2014; Sample & Michel, 2014; Tavakoli, 2016; Vercellotti, 2012, 2015). Because spoken language in general and learner language data in particular require a determined set of rules as to how to segment, Foster et al. (2000) give detailed explanations where exactly to put the AS boundary in a variety of examples. For this reason Ellis and Barkhuizen (2005) strongly recommend that researchers consult Foster et al. (2000) before segmenting any spoken language data.

The AS-unit is considered the most reliable unit of choice for the data of the current study as well since the interviews were conducted orally without rehearsal time and included spoken language only. In addition, the speakers were L2 speakers of English and were expected to show sentence fragments. The AS-unit is the unit that not only includes all data in the sample in the analysis instead of abandoning those segments that are difficult to capture. It also provides a fairly rigorous set of practical considerations how segmenting is done, which add to its reliability and make results of a data analysis comparable to a great extent.

In addition to the challenges how to sequence productive language data, each CAF dimension bears difficulties as to how it can be measured reliably, validly, and feasibly. The following sections show how complexity, accuracy, and fluency can be measured as individual dimensions of language performance.

2.1.4.2 Complexity

The one most prominent and reoccurring measure of syntactic complexity is a ratio of subordination (see overview in Norris & Ortega, 2009). By linking simple sentences through subordination, language becomes more complex. The reason for including subordination as the measure of syntactic complexity is based on the idea that subordination is the most important indicator of syntactic complexity.

In all measures, clauses function as the main determinant in measuring syntactic complexity. For example, Larsen-Freeman (2006) measured syntactical complexity in the average number of clauses per T-unit. Similarly, Ellis and Yuan (2004) measured syntactic complexity in the ratio of clauses to T-units. Other studies have also applied the number of dependent clauses per number of total clauses, or the number of dependent clauses per T-unit (see Larsen-Freeman, 2009).

Apart from using different base units such as the T-unit and AS-unit, there is some variation of what is considered a clause. According to Hunt (1965), a clause is a "structure with a subject and a finite word (a verb with a tense marker)" (p. 15). However, as Bulté and Housen (2012) remark, this definition excludes complex structures with non-finite clauses. Foster et al. (2000) redefined the clause for the purpose of consistent research to include any verbal element, finite or non-finite.

In addition to a measure based on subordination, few studies included other morpho-syntactical features in their assessment of complexity. Complexity in terms of morphological measures, however, has not been in

use widely (Bulté & Housen, 2012).[9] As one of few studies, Ellis and Yuan (2004), for example, also included syntactic variety in terms of the number of different verb forms used: "Grammatical verb forms included tense (e.g., simple past, past continuous), modality (e.g. *should, have to*), and voice (e.g., passive voice in the past)" (p. 71). As was explained in section 2.1.1, such measures are geared to capture the development of a specific form over time and are not intended to represent a global picture of complexity in a speaker's language. Picking up Pallotti's (2015) aforementioned call for a "simple view" on complexity again, "[a]ll measures representing syntactic complexity in terms of structures' difficulty, sophistication and acquisitional timing are at odds with the simple view proposed here" (p. 125).

Another perspective is suggested by Norris and Ortega (2009), when they argue that at least three complexity measures need to be measured in order to capture global complexity, phrasal complexity, and complexity by subordination. They reason that language can be elaborated at three different syntactic levels. Accordingly, global complexity is measured in length of unit, for example in words per AS-unit. Global complexity is therefore based on the assumption that with length, also the complexity of an utterance increases. Phrasal complexity has been suggested as becoming a relevant feature of syntactic complexity "at the most advanced levels of language development and maturity" (Norris & Ortega, 2009, p. 563). Findings of more phrasal complexity at the very advanced level were found for example in Biber (2006) for academic texts of mature L1 English speakers. At that level, increased phrasal complexity may add to the amount of subordination rather than that the amount of subordination increases (Norris & Ortega, 2009).

Subordination has become the most consistently used measure of syntactic complexity. However, Ellis and Barkhuizen (2005, p. 155) point out that complexity by subordination can only be measured validly if the learners have acquired subordination constructions to some degree. Subordination as a measure of syntactic complexity has therefore shown to be most valuable at intermediate and upper-intermediate levels (Norris & Ortega, 2009, p. 564).

Lexical diversity, as the other dimension of complexity, has often been calculated in the field of linguistic research using the ratio of different words to total words, the Type-Token Ratio (TTR), or measures based on it.

9 Publications after the completion of the current dissertation indicate that morphology may be incorporated more in future studies (e.g. Kuiken, Vedder, Housen, & De Clerq, 2019). A morphological measure is particularly suggested for studies on morphologically rich languages such as French, on the development of a language over time, or comparisons between different L2s (De Clercq & Housen, 2019).

The TTR, however widely in use, has not been considered a valid measure to include texts, written as well as spoken, of differing lengths because language samples with more tokens will result in lower type-token ratio values and vice versa. As McKee, Malvern, and Richards (2000) explain, "[t]he reason for this is very simple – as longer and longer samples of language are produced, more and more of the active vocabulary is likely to be included and the available pool of new word types that can be introduced steadily diminishes" (p. 324). Thus, the longer the text is, the predictably lower the TTR. This problem in determining lexical diversity has been addressed frequently in more recent years (R. Ellis & Barkhuizen, 2005; Jarvis & Daller, 2013; Larsen-Freeman, 2006; McCarthy & Jarvis, 2007; McKee et al., 2000; Skehan, 2003).

Solutions to solve the problem were, for example, measuring vocabulary complexity by a type-token ratio calculated in word types per square root of two times the words (R. Ellis & Barkhuizen, 2005; Larsen-Freeman, 2006). More comprehensive validation studies, however, revealed that lexical diversity can be calculated much more reliably in a measure D vocabulary diversity, of which the most common one is vocd[10] (Malvern, Richards, Chipere, & Durán, 2004; McCarthy & Jarvis, 2007, 2010; McKee et al., 2000). Vocd was developed to overcome the high dependency on text length of the type-token ratio (Malvern & Richards, 2002).

D is based on comparing the mathematical model with the empirical data in a transcript and "provides a new measure of vocabulary diversity" (McKee et al., 2000, p. 324). Vocd is calculated based on several random text samplings. McCarthy and Jarvis (2010) explain the calculation as follows:

> The approach begins its calculation by taking from the text 100 random samples of 35 tokens. The TTR for each of these samples is calculated, and the mean TTR is stored. The same procedure is then repeated for samples from 36 to 50 tokens. An empirical TTR curve is then created from the means of each of these samples. (p. 383)

Lexical diversity as calculated in vocd has been considered a "generally acceptable measure" (Skehan, 2009, p. 514) to measure the lexical range. Vocd has been in use as a measure of lexical diversity in a number of different studies in language acquisition: for example, second language acquisition in twins (e.g. Lowie, Verspoor, & van Dijk, 2018), child-directed speech and first language acquisition (e.g. Rowe, 2008), first and second language fluency (e.g. Skehan, Foster, & Shum, 2016), stuttering in children (e.g. Silverman & Ratner, 2002), or as part of the CAF dimensions and their

10 Variant spellings can be found, also vocD, VocD, VOCD with no difference in meaning.

development (e.g. Kormos & Trebits, 2012; Pallotti, 2015; Skehan, 2009; Skehan et al., 2016; Vercellotti, 2015), to name but a few. McCarthy and Jarvis (2007) also consider the vocd output as "a relatively robust indicator of the aggregate probabilities of word occurrences in a text" (p. 459), even though they have pointed out that vocd may also be effected by text length (McCarthy & Jarvis, 2007, 2010). When vocd was applied, those texts that had between 100 to 400 tokens showed to be comparable (McCarthy & Jarvis, 2007).

To conclude, subordination is the predominant measure applied to indicate syntactical complexity. Even though the participants' language levels in the present study was not known, there was reason to assume they would be at least at a low-intermediate level, at which some subordination may be expected to have been acquired: All of them were teaching regular English classes at the time of research and volunteered in taking part in the study. Therefore, subordination per AS-unit and time was used to measure syntactic complexity. A subordination measure was also expected to be the measure that could capture complexity in comparably low as well as in comparably high performers, should there be a noticeable range in the speaking performances.

In addition to using a subordination measure, a length measure – words per AS-unit – was included, as was proposed by Norris and Ortega (2009), because the study at hand was interested in capturing the speakers' global complexity. Adding a length measure also guaranteed to capture some complexity in case there was going to be little subordination in the data of the present study.

In order to measure lexical diversity, a variety of measures have been in use in research, of which the type-token ratio has been widely criticized for being greatly affected by text lengths. Lexical D has emerged as the most widely used one, claiming the status of an "industry standard", as McCarthy and Jarvis (2007, p. 461) state. Vocd has been suggested as a measure to capture lexical diversity that is not insensitive to text length, but one that is considered a reasonably reliable measure among other measures such as HD-D and the measure of textual lexical diversity MTLD (McCarthy & Jarvis, 2010).

It was reported that texts of 100 to 400 tokens "might be suitably compared" (McCarthy & Jarvis, 2007, p. 481) using vocd. Vocd is also a measure that has a "rich history" (McCarthy, 2017) and a measure whose frequent use in research makes it comparable and valuable. Lexical D as measured in vocd was used to measure lexical diversity in the present interview study.

2.1.4.3 Accuracy

As the term accuracy suggests, measures predominantly focus on the amount of error-free language and rarely on counting individual errors. For example, accuracy is measured as the proportion of error-free T-units to total number of T-units (Larsen-Freeman, 2006), or in other studies also the number of error-free T-units and errors per T-unit (Larsen-Freeman, 2009). Most studies calculating accuracy in the CAF framework report the amount of error-free language per given unit. Ellis and Yuan (2004), for example, calculated the percentage of error-free clauses by excluding clauses containing errors in syntax, morphology, and lexical choice. Clauses containing errors were not counted as error-free but were included in the total count of clauses to arrive at the percentage of error-free clauses. Ellis and Yuan (2004) also included a specific measure and calculated the percentage of correct verb forms "in terms of tense, aspect, modality, and subject-verb agreement" (p. 72). Thus, they use what is referred to as global accuracy and local accuracy. Foster and Wigglesworth (2016) distinguish local and global accuracy measures and their applicability as follows: "[L]ocal measures of accuracy are most valid in circumstances where the focus is on how development of a grammatical morpheme responds to particular treatments. Global measures, by contrast, examine the text or transcript in its entirety" (p. 102). The distinction between global accuracy and local accuracy, for example the correct use of plural forms, is a similar distinction discussed relating to syntactic complexity in sections 2.1.1 and 2.1.4.2.

Specific forms, such as past tense verb forms, can therefore serve to indicate development in the acquisition of a particular morpho-syntactic form. Error-freeness in clauses on the other hand represents the accuracy of a sample as a whole, meaning globally.

In a recent study, Foster and Wigglesworth (2016) argue that errors can differ in their severeness, which is not accounted for in measures of error-freeness. A morphological error that is barely noticeable weighs as much as several errors that render the particular clause unintelligible. For this reason, Foster and Wigglesworth (2016) put forward an additional approach to measuring accuracy – one that integrates gravity of error. They suggest categorizing errors into three levels to calculate a "weighted clause ratio" (Foster & Wigglesworth, 2016). The errors in a clause are graded "minor", "serious", and "very serious", levels 1 through 3 respectively (Foster & Wigglesworth, 2016, p. 106). The clauses are then multiplied by factors 0.80, 0.50, or 0.10 according to the gravity of the errors, or factor 1 if the clause is entirely correct. While categorizing errors into *entirely accurate* versus *compromising comprehensibility* can be expected to be reliably decided by the raters, the level 1 and 2 middle range of errors, however, is most prone to variability in the rating (Foster & Wigglesworth, 2016).

As discussed in section 2.1.2, adequacy and appropriateness are considered part of a separate level in language use that has not been incorporated in the CAF framework because concepts of adequacy and appropriateness are not necessarily grounded in structural accuracy. A weighted error scale, such as the one introduced by Foster and Wigglesworth (2016), may serve a research purpose well if different types of errors, which represent particular stages of language development, or a change in accuracy over time are the focus of research. For the present study, however, accuracy needed to be measured on a global scale to obtain an overall score of accuracy for each participant, regardless of the types of error and their development over time.

Another reason to recommend error-freeness as a measure is mentioned in Ellis and Barkhuizen (2005). They propose a global measure of accuracy such as the proportion of error-free clauses, because if learners avoid particular forms, their accuracy would be misrepresented in the specific measures. Likewise, an account of accuracy based on specific measures of accuracy is geared toward research on a targeted structure, but it is considered less fit to capture overall accuracy performance (Vercellotti, 2012).

With respect to the choice of the unit, on which to base the accuracy measure, Foster and Wigglesworth (2016) comment that "the best tool in the measurement kit is currently the error-free clause, because it combines a reliably defined and valid unit with a finer-grained analysis than offered by a whole T-unit or AS-unit analysis" (p. 104). However, supporting comparative studies in the measures would be needed to justify using a clause-based measure only. Instead, an additional measure based on AS-unit may add to capture accuracy more extensively in those places where a clause-based count might exclude cases.

Concluding, measuring accuracy in terms of error-free clauses, when accuracy is to be accounted for globally, has shown to be the most useful measure to date, even though inconsistencies in judging the error-freeness of a clause cannot be ruled out entirely. Since the present study aims to obtain a global indication of the interviewees' accuracy as part of an over-all performance, two measures were used – the percentage of error-free clauses and the ratio of error-free clauses to AS-unit.

2.1.4.4 Fluency

Typically, measures for all three subdimensions are involved when oral performance is measured in terms of fluency (Tavakoli & Skehan, 2005): breakdown fluency as in number and length of pauses, speed fluency as in speech rate and density per time unit, and repair fluency as in false starts and repetitions per unit. Each of those aspects can be acoustically measured.

Breakdown fluency is measured in the number of silent pauses per time unit, the number of filled pauses per time unit, and the mean length of

silent pauses (Bosker et al., 2012). There is consent that pauses in terms of silences affect measured as well as perceived fluency and pauses therefore play a core role in measuring fluency. However, there has been a variety of differing thresholds regarding how long a pause needs to be in order to count as a silent pause in measuring fluency. Goldman-Eisler (1968) was one of the earliest ones to propose that the length threshold of a silent pause be reasonably set at 250 ms to be considered a hesitation. Because studies investigating fluency varied in the thresholds of silent pauses, several decades later, N. H. de Jong and Bosker (2013) examined which silent pause threshold would correlate most with L2 proficiency as measured in a productive vocabulary test. They argued that the higher the correlations between a specific threshold and L2 proficiency, measured in vocabulary knowledge and rated fluency, the more support there would be for that specific threshold (N. H. de Jong & Bosker, 2013, p. 17). Indeed they found the highest correlation when the silent pause threshold was between 250 and 300 ms. N. H. de Jong and Bosker (2013) therefore suggest working with a 250 ms threshold for silent pauses when dealing with L2 data. Pauses shorter than 250 ms are considered micro-pauses that are not included in the hesitation phenomena (Riggenbach, 1991).

Speed fluency is calculated in the number of syllables per time unit. Since pausing affects the actual speaking time, some studies use speaking time instead of total time as the base for all fluency measures (e.g. Bosker et al., 2012). That way speed and breakdown measures will not be confounded as speaking time, called phonation time, is the time minus the silent pauses. Another measure to indicate fluency is the number of words in a unit. Larsen-Freeman (2006), for example, used the average number of words per T-unit to measure fluency in her study of Chinese English speakers. Some debate revolved around whether a measure of words per syntactic unit, such as the T-unit, should be a complexity or fluency measure (Norris & Ortega, 2009). On the one hand, Norris and Ortega (2009) suggest that a length-based measurement based on a syntactic unit ought not be considered a measure of fluency but rather of complexity. Likewise, they also argue against using time units in a complexity measure, because complexity and fluency would then be conflated. Convincingly, speed fluency is rather based on time than on a syntactic unit, which has been accounted for in the present study.

Repair fluency is measured in repair phenomena, as in the number of repetitions and the number of self-corrections per unit (Bosker et al., 2012). The repair measures' reliability is considerable, as there is only little risk of error in the measure itself: repair phenomena can be included in the data transcriptions and counted.

Because fluency is multidimensional and includes a variety of measures for each of its sub-dimensions, it takes a painstaking effort to calculate all measures. Software applications such as PRAAT (Boersma &

Weenink, 2005) have been developed, which automatically detects pauses in recordings. Scripts, for example the Praat Script Syllable Nuclei (N. H. de Jong & Wempe, 2009), help the application to detect syllables and calculate a number of speech rate measures. Since the application tracks all sounds, complementary transcripts are needed to measure all those fluency measures that are qualitative in nature, such as repair phenomena and filled pauses.

Fluency needs to be measured in a number of indices, if large parts of its multidimensionality, specifically in spoken language, are to be captured as part of a linguistic performance analysis. Fluency is not described in one subdimension only but all three, speed, breakdown, and repair, form the particular fluency of a speaker. Therefore, fluency in the current study measures all three subdimensions – speed fluency, breakdown fluency, and repair fluency. Each of the subdimensions is measured in a combination of numerous measures. The measures applied in the current study are listed in Table 2 in section 3.2.2.2.4 of the empirical study.

2.1.5 Factors Influencing CAF

The previous section has shown that CAF measures are numerous and may differ in the studies leading to results that are difficult to compare. Apart from some inconsistencies in measures applied in studies, which may result in differing findings, CAF performance can differ depending on the task as well as between native and non-native speakers. In an attempt to answer whether and how performance on CAF may be influenced by the quality of the task, some studies investigated the relationship between task and performance (N. de Jong & Vercellotti, 2015; Foster & Tavakoli, 2009; Michel et al., 2007). Foster and Tavakoli (2009), for example, found that tasks asking for narratives that included background and foreground events increased syntactic complexity: connecting background and foreground events is done through subordinate clauses. In addition, Foster and Tavakoli (2009) not only looked at how different tasks influenced the outcomes of their subjects in terms of CAF, but also if native and non-native speakers performed differently. They found that subordination is used by both native and L2 speakers to combine background and foreground narratives.

Fluency, however, was affected by how tightly-structured the narrative task was: tightness in the story-telling design, which meant that the order of events in the story was fixed and could not be changed, helped the non-native speakers increase fluency, whereas it had no effect on the native speakers' fluency. In addition, Foster and Tavakoli (2009) observed a difference as to where L2 and native speakers pause. L2 speakers' pausing was observed more often in the middle of clauses while native speakers' pauses occurred at syntactic boundaries but significantly less often mid-clause. These results

indicate that the native speakers in the study planned their speaking along syntactic boundaries and also retrieved their language in syntactic chunks, whereas the L2 speakers did not show to do so and interrupted their flow of speech mid-clause.

Because language learners are expected to encounter difficulties in planning and retrieving language while speaking, planning time emerged into the focus of some studies to investigate whether speakers performed differently when they were given planning time before doing the task. Several studies found that pre-task planning positively affected fluency, complexity, and accuracy (Bamanger & Gashan, 2015; Foster & Skehan, 1996; Kawauchi, 2005; Tavakoli & Skehan, 2005; Wigglesworth, 2001). However, the results on the effects of planning time are not entirely consistent as some studies found no positive effect on some CAF measures (Gilabert, 2007; Yuan & Ellis, 2003). Bamanger and Gashan (2015) suggest that in the studies in which planning time did not enhance the outcomes, the participants might not have been familiar with task planning to make efficient use of the planning time. Rehearsal, as another factor possibly influencing the CAF outcomes, was examined by R. Ellis (2009). He found that rehearsal in completing a particular task resulted in greater fluency and complexity. However, the rehearsal effect did not translate to other tasks and left the results non-generalizable.

How task type can relate to speakers' performance was shown in Ferrari (2012). One of her results indicated that the participants initially performed highly complex sentences on a telephone call task, but gradually decreased their complexity over the following years. This result also illustrates that high complexity is not necessarily considered adequate in all situations: the native speakers in the study used much less complex sentences to start a telephone conversation. De Jong and Vercellotti (2015) show that different topics may elicit different forms as well, and some topics elicit more lexically diverse language than others.

Concluding, a reasonable point is made when Vercellotti (2015) emphasizes that most studies in the CAF frame are based on aggregated results of compared group means, not on within-individual findings, and are therefore limited in what they can suggest in terms of how the dimensions complexity, accuracy, and fluency interact within an individual. She considers within-individual insights necessary in order to claim effects on CAF performance in the individual speaker. Between-group mean comparisons might also account for the inconclusiveness of the studies examining the effect of planning time on performance, as for example Yuan and Ellis's (2003) study compared mean scores of each planning group.

The present study looks at individual performances and bases the relationships between complexity, accuracy, and fluency on individual scores within the group instead of means between groups. While the

study cannot be considered a within-subject study in Vercellotti's terms, as it does not focus on individual development over time, it can give an account of cross-sectional individual performance on complexity, accuracy, and fluency and their relationship in the L2 production of the speakers studied. The following section outlines the research positions regarding the relationships between the dimensions.

2.1.6 Trade-Off Effects

A core issue in the understanding of L2 language production in the CAF framework is how complexity, accuracy, and fluency relate to one another. Whether or not there is a competition between different aspects of language performance while speaking a foreign language has been a matter of debate in various areas of second language research. For one, observed interrelationships between aspects of language performance could support an understanding of procedural processes taking place while speaking. For another, insights into the relationships of the complexity, accuracy, and fluency dimensions could influence approaches to foreign language teaching, for example form-focused activities versus fluency-focused tasks in foreign language instruction as described in Thornbury (2000), for example.

Several hypotheses have been stated with regard to the relationship between the dimensions, and a large body of research is still being conducted aiming to shed more light on their interrelationships. Skehan (2009) proposes trade-off effects between complexity, accuracy, and fluency, known as the Trade-Off Hypothesis or earlier as Limited Capacity Model (Skehan, 1998). The Limited Capacity Hypothesis suggests that complexity, accuracy, and fluency compete in language production due to limited attentional capacity and working memory. According to this hypothesis, all three dimensions compete and cannot be attended to simultaneously when speakers concentrate on meaning. If speakers focus on form, there is another, secondary contrast between accuracy and complexity, indicating that controlled accuracy trades off with complexity. Several of Skehan's original articles (e.g. Foster & Skehan, 1996; Skehan, 2009, 2015) have been republished as a book (2018), which may indicate the ongoing relevance of the topic.

The Cognition Hypothesis (Robinson, 2003, 2011) presumes competition among the dimensions in language performance as well, but it suggests a different trade-off effect: tasks may promote either fluency, or complexity *and* accuracy. If the tasks demand so, complexity and accuracy will be linked to one another more closely than fluency. The Dynamic Systems Theory or Complexity Theory (e.g. de Bot, Lowie, & Verspoor, 2007b, 2007a; Larsen-Freeman, 2009) add another perspective and claim that trade-off effects

between the dimensions may be observed, but are not causal, linear, or mutually exclusive.

Study findings show a non-conclusive picture on the issue of trade-off effects between complexity, accuracy, and fluency. When Ellis and Barkhuizen (2005) examined how different tasks influenced complexity, accuracy, and fluency, they found that tasks could lead to one dimension outperforming the others. In her longitudinal study on within-subject performances, Vercellotti (2015) found no trade-off effects between complexity, accuracy, and fluency but instead, correlations between all three dimensions. She also points out that "in many of the between-group designs which lead to conclusions of trade-off effects, the groups represent different CAF-focused performances" (Vercellotti, 2015, p. 2). Trade-off effects between fluency and accuracy have been reported in Yuan and Ellis (2003), Michel et al. (2007) and Ahmadian and Tavakoli (2011). On the other hand, grammatical complexity came at the expense of fluency for instance in Bygate's study (2001). Skehan and Foster (1997) as well as Ferrari (2012) showed that complexity and accuracy competed in their studies.

Studies taking lexical diversity into account found that either lexical variety and accuracy traded off (e.g. Yuan & Ellis, 2003), lexical variety and accuracy increased jointly (e.g. Robinson, 1995), or lexical variety was positively correlated with accuracy, but negatively with grammatical complexity (e.g. Skehan, 2009).

As Vercellotti (2015, p. 4) notes, much of the research on trade-off effects is based on cross-sectional between-group means (e.g. Bygate, 2001; Skehan & Foster, 1997; Yuan & Ellis, 2003). Means of different groups, for example three planning types in Yuan and Ellis (2003), were compared and showed trade-off effects. However, there was no trade-off effect within each planning group. Similarly, Skehan and Foster's (1997) study found trade-off effects between the means of different task groups, but not within each task group.

Observed findings in the field of the relationships between the dimensions complexity, accuracy, and fluency are contradicting and non-conclusive. Results may have been affected by a variety of differing elements in the research designs such as different CAF measures, tasks, or task conditions, as well as group versus individual comparisons, all of which make the comparability of the findings difficult. However, whereas there is virtually no disagreement that L2 speakers can focus on a particular dimension, the questions of whether they inevitably have to is much less agreed on (Vercellotti, 2015). Therefore, observations of trade-off effects, their existence and absence alike, need to be considered within their relationship to the respective research design.

2.1.7 Summary and Discussion

Complexity, accuracy, and fluency have shown to be viable dimensions of language production in second language acquisition research. They contribute to overall language proficiency, yet performance and proficiency are not synonymous.

While the nature of interrelationships between the CAF dimensions remains complicated, there seems to be agreement that "all three must be considered if any general claims about learners' L2 performance and proficiency are to be made" (Housen et al. 2012a, p. 3). The present study therefore focuses on the teachers' English language performance according to the CAF framework. The participant teachers in the study at hand embody dual roles: one as second language speakers of the target language and another one as teachers and model target language providers to their students.

Part of the reason for diverging research results in the field of language performance is that measures across studies are rarely the same – a situation evoking a claim that has been proposed for the past years: "A significant progress in the field would thus be the identification of a limited set of standardized measures to be used across studies" (Pallotti 2009, p. 17). Progress has been made since in research in clarifying definitions and shaping the measures as for example in Inoue (2010), Housen et al. (2012b), de Jong and Bosker (2013), de Jong et al. (2015), Pallotti (2015), Tavakoli (2016), Foster and Wigglesworth (2016), and Wright and Tavakoli (2016).

For the current study, a task design was necessary that could elicit as much of the speakers' potentials in terms of complexity, accuracy, and fluency. In addition, since the results were also expected to inform on the relationships among complexity, accuracy, and fluency in language production, a within-group analysis was needed. With a within-group design, performances could be compared based on the individual performances in the group while keeping the task the same. Further, a model needed to be developed that could calculate all the relevant measures of CAF and transform them into an index that could further be used in computing the relationships between the teachers' language performance and their students' outcomes in receptive vocabulary and grammar.

If linguistic performance as part of a more general language proficiency is not only studied in its own right, but also as a possible factor influencing students in their second language acquisition, the question needs to be examined how linguistic performance could play a role, which mechanisms may be at play between language input and language acquisition, and what research has found to date that would suggest some relevance of the teachers' linguistic performance for their students' second language acquisition. The following sections focus on input as it is defined, discussed, and studied.

2.2 Second Language Acquisition and Linguistic Input

Ortega (2015) identifies five central fields in second language acquisition research:

> (a) the nature of second language knowledge and language cognition, (b) the nature of interlanguage development, and the contributions of (c) knowledge of the first language (L1), (d) the linguistic environment, and (e) instruction. (p. 245)

The contribution of the linguistic environment is the area in which the study of teacher language performance and its link to children's second language acquisition may be situated best: The teachers' language performance and its possible effect on the children's second language acquisition is a main focus. The teachers' productive L2 performance as well as the students' receptive L2 development also inform about second language knowledge, which is named as the first central field in second language acquisition research, (a) in the list above.

In order to analyze the contribution of the linguistic environment, it is necessary to discuss the role that linguistic input is assigned in second language acquisition. Input ranges among numerous internal and external factors that are subject to research as to whether and how they influence second language acquisition. Among many more, these factors include for example aptitude, phonological awareness, age of onset, duration of instruction, socio-economic status, or amount and quality of input, all of which are acknowledged in a wide range of second language acquisition research (e.g. Cook, 2008; Cook & Singleton, 2014; Dörnyei, 2014; Doughty & Long, 2003; R. Ellis, 1997, 2015; R. Ellis & Shintani, 2014; Gass, 2013; Gass & Selinker, 2001, 2008; Larsen-Freeman & Long, 1991; Myles, Mitchell, & Marsden, 2013; Ortega, 2011, 2009; Saville-Troike, 2016; Sharwood Smith, 1994). Each of those factors is subject of a growing body of research intending to advance the understanding of how internal factors, for example aptitude, and external factors, for instance the amount of target language exposure, act in second language acquisition processes.[11]

While acknowledging that all of the aforementioned factors may affect language acquisition, the present study focuses on teachers' linguistic performance and how it relates to students' language acquisition. Connecting teachers' performance and students' outcomes draws on theoretical considerations that link input to language acquisition. The

11 For a summary of the influencing factors on SLA, see Kersten (2019).

notion of linguistic input, however, is not based on an unequivocal understanding across disciplines and studies. The following illuminates some of the common as well as divergent understandings of input and discusses the concept against the background of the current studies.

2.2.1 Defining Input

Linguistic input is one of the external factors in language acquisition. It is referred to as "the samples of language to which a learner is exposed" (Ellis, 1997, p. 5). Gass (1997) proposes the same understanding of input and adds the mode of exposure: It is "the language to which a learner is exposed either orally or visually (i.e., signed languages or printed matter)" (p. 28).

In this sense, input is a prerequisite of language learning and acquisition: "If learners do not receive exposure to the target language they cannot acquire it" (Ellis, 2005, p. 216). Language input has to be provided to the learners in order for them to be able to develop in a language. Carroll (2001) introduces using the term *stimuli* instead of *input* to refer to terminology of psychology of learning. She defines *stimuli* as the "observable instantiations of the second language" (Carroll, 2001, p. 8). Larsen-Freeman (2015) proposes "ambient language" instead of the term *input*, emphasizing an understanding against a perception of input as a static concept. Because the term *input* is commonly used in large parts of second language research as well as first and bilingual language acquisition research, the term *input* will further be used in the current study, yet in essence based on the same definition proposed by Carroll for *stimuli* while acknowledging Larsen-Freeman's stand.

Main ideas on how linguistic input may relate to second language acquisition are presented and discussed in the following section. Input plays a role in all second language acquisition theories, but with slightly different weights and consequences: Universal Grammar (UG) Theory gives input comparably little weight, whereas usage-based approaches to second language acquisition assign a central role to input, to name those two theoretical approaches that can be considered being at the opposite ends of a continuum of how relevant input is considered to be for second language acquisition (Ortega, 2015). According to theories based on UG, input is needed only to trigger principles and parameters of a particular structure (White, 2015). In usage-based approaches, on the other hand, language learning is based on "statistical estimations" (N. Ellis & Wulff, 2015, p. 86). In those approaches, learners need input to associate constructions. Input is therefore needed in large amounts, as it is believed to drive the language learning process.

Research on how input affects language acquisition, first or second, may consider input in terms of its linguistic properties, amount, duration, or even methodological strategies. How particular aspects in the linguistic input interrelate with language acquisition, is a question still not fully answered. The working understanding in a usage-based approach to second language acquisition is that "'rules of language, at all levels of analysis (from phonology, through syntax, to discourse), are structural regularities that emerge from learners' lifetime analysis of the distributional characteristics of the language input" (N. Ellis, 2002, p. 144). In this sense, the learner constantly analyzes the distributional properties of the language input, forms rules based on those, tests and retests hypotheses about the linguistic system of that language, or put differently: "Learners have to *figure* language out" (N. Ellis, 2002, p. 144). As one of the more recent papers discussing UG and SLA, Rankin and Unsworth (2016, p. 564) stress that input messiness and ambiguity are genuine features of the generative acceptance of Poverty of Stimulus (POS), while they also call for more robust research on input to understand its role in second language acquisition.

The internal mechanisms, meaning how learners process input rather than the input itself, is considered central in processing theories such as Processability Theory, coined to Pienemann (1998), or Input Processing Theory, suggested by VanPatten (1996). How exactly language processing is executed, can therefore be considered a subject of its own in theories on language acquisition.

It is the understanding of input as potential linguistic data, which is at the scope of the current study. This understanding focuses on the linguistic properties of the language providers' spoken language and considers methodological teaching strategies and how they might be affected by the overall target language proficiency of the teachers as a separate field of research, albeit interrelated and no less important. As the present study focuses on linguistic performance as a part of the over-all language proficiency, an extensive evaluation of classroom input and teaching is not within the scope of the project at hand. As such, input neither includes any assumptions about what exactly the learner hears, identifies, or even processes, nor methodological approaches to how to deliver linguistic input.

2.2.2 Modeling Input to Output

Even though language processing and its specifics are not at the focus of the present thesis, the question how input may move to output in second language acquisition is of some relevance for the motivational background of the present study. Two models of second language acquisition, one that has been referred to frequently for several decades, Gass (1988), and

another more recently introduced model by Leow (2015), are explained in the following to illustrate how input may transform to output in second language development.

Gass's illustration of input in second language acquisition was brought forth in 1988 and republished with only insignificant modification several times thereafter (e.g. Gass, 2013; Gass & Selinker, 2001). The model represents a foundation for parts of the discussion of the present study results as it introduces specific terms and ideas reviewed in the following.

In the five-stage model of second language acquisition, input is the language data that "filters through to the learner" (Gass, 1997, p. 4). Input is represented in the model by a dotted line indicating that some of the input might trickle through, some might not. Input is located on top of all the stages that follow. It is present before any second language acquisition process can kick off and the five stages of language acquisition – apperception, comprehension, intake, integration, and output – come into place. The five stages indicate the following (Gass, 2018, pp. 3–8): Input is apperceived when bits of language in the input are noticed and recognized as new features. Comprehension takes place if the recipient understands. Understanding can relate to parts of the input, such as particular forms, or the message in general. At the third stage, which is referred to as intake, input data is processed. Generalization of data takes place here, and memory is set against prior knowledge. Integration is the stage where storage and a second language grammar develop. Finally, output is the fifth stage in Gass's second language acquisition model, at which the transformation into output – or L2 language production – materializes.

According to Gass's model, input can take four possible routes: The first is where the learner's hypotheses about the language features are either confirmed or rejected, depending on whether the input delivers supporting or rejecting data for a hypothesis. Once a hypothesis is confirmed, input data will move to be integrated. The second route is similar, except that in this case information is already incorporated in the grammar. Input data may then strengthen the existing hypotheses. A third possibility is that input is put into storage. This may happen when some understanding has taken place, but when there is still uncertainty of how to integrate the particular feature into the learner's grammar. There may have been insufficient information in the input data that would otherwise motivate the learner to reject or to confirm a hypothesis. Gass (2018) proposes that vocabulary and smaller chunks of language are more likely to be moved to storage than longer syntactic strings because it is "more difficult to hold large bits of language in memory for a long period of time" (p. 7). Finally, the fourth path in Gass's model is the exit, at which learners "make no use of the input" (p. 7).

In Gass's model, several factors can influence the route input takes in second language acquisition: time, frequency, affect, prior knowledge, salience, and attention. Time and frequency are the ones most relevant to understanding the results of the study at hand and are therefore explained here. Time pressure is considered to be particularly relevant at early stages of second language acquisition for the following reason: "[M]uch of the input is difficult to separate into words or phrases or other units that may be manageable by the learner" (Gass, 2018, p. 17). Therefore, input that is broken down into smaller chunks by pausing and a slower speech rate is considered to support language acquisition – an aspect to be discussed in light of the results of the present study (section 3.4.2.3). Frequency refers to the rate at which an item occurs in the input. Not only features that are highly frequent in the input may be noticed more easily, but also unusual and rare features, even though the latter phenomenon is more relevant at advanced stages of second language acquisition (Gass, 2018, p. 17).

Leow's (2015, 2019) model of the L2 learning process based on instructed language learning also suggests stages from input to output. This model illustrates input processing in a circular fashion with an opening for input and an exit for the product. Similar to Gass's model, it assumes intake is a stage between input and L2 knowledge, alternatingly passing what is called "process" and "product" (Leow, 2015, p. 242). The model distinguishes between three forms of intake: attended intake, detected intake, and noticed intake. This distinction is made in order to account for different stages of attention and processing. In Leow's model, attention is assigned a central role, as it is considered the prerequisite of processing and intake. There are similarities to the earlier model by Gass, in particular in the notion of intake as a process taking place en route from receiving language input and producing language. An additional feature in Leow's model is that output may also loop to function as input – an aspect essential in the output hypothesis mentioned below (section 2.2.3). As attention and output as part of language processing will not be at the focus of the current study, a less fine-grained model such as Gass's suitably serves the purpose as a referential model to draw on.

2.2.3 Theoretical Hypotheses of Input Relevance

Input not only plays a role in models of second language acquisition, such as Gass's (1988, 2013) and Leow's (2015, 2019) as discussed in section 2.2.2, but also in hypotheses on what may assist second language acquisition. Several theoretical hypotheses make predictions as to how linguistic input may function and relate to second language acquisition. Some theoretical hypotheses may then result in teaching strategies that are believed to aid

students in processing linguistic input and taking in the target language features. An overview of those hypotheses and theoretical considerations that benefit a discussion of the current results in the context of instructed language learning is discussed in the following.[12]

For example, the Incidental Learning Hypothesis (Schmidt, 1990, 1994a) argues from a perspective that looks at whether language development takes place incidentally or intentionally. It is concerned with how much of a planned process underlies language acquisition. According to the Incidental Learning Hypothesis, a large part of second language acquisition takes place incidentally. It is "learning of one thing (e.g., grammar) when the learner's primary objective is to do something else (e.g., communicate)" (Schmidt, 1994, p. 16). This hypothesis "acknowledges that much of the learning that takes place is associative in nature" (R. Ellis & Shintani, 2014, p. 175). Learners are able to plot form to meaning if they are exposed to a number of linguistic forms. Incidental learning has been particularly emphasized in vocabulary development (e.g. Laufer & Hulstijn, 2001; Loewen & Sato, 2018; Newton, 2013).

The Frequency Hypothesis proposes that frequency of particular linguistic forms in the input determines the sequence of the acquisition of those very forms (N. Ellis, 2002). According to this hypothesis, more frequent features in the language input would be acquired sooner. Some objections came from observations with speakers whose first language did not show some of the highly frequent features in English: Japanese, for example, does not have definite and indefinite articles, which makes it difficult for L1 Japanese speakers to acquire articles in English, despite their high frequency in the English input, as Ellis and Shintani (2014) report. Even though the Frequency Hypothesis may not be able to provide a comprehensive explanation of second language acquisition, as Ellis and Shintani (2014, p. 176) also point out, it has influenced current thinking in second language acquisition and input effects, and some form of frequency is referred to in most theoretical considerations of input effects.

The Input Hypothesis, a part of the Monitor Theory (Krashen, 1981, 1985), implies that language would automatically be acquired through comprehensible input that is slightly above the learner's target language level, referred to as i+1, 1 indicating the next level. It then follows that the teacher's role in language learning is that of a provider of sufficient comprehensible input, meaning simplified input. This view was much criticized for several reasons, one of which was concerned with defining *comprehensible*, another of which did not consider such an automaticity to be supported by research findings (e.g. Carroll, 2001; White, 1987). To the contrary, studies showed

12 For a further discussion on language teaching approaches, see J. C. Richards and Rodgers (2001).

that despite what was considered comprehensible input, children as well as adults differed greatly in their second language development even if they were the subjects of similar language input, the reason being that learners can comprehend language in a top-down manner by using clues in the context only (R. Ellis & Shintani, 2014, p. 177). Carroll (2001) comments that "Krashen got it backwards!" (p. 9), which means that knowledge of grammatical properties is necessary to be able to parse a second language rather than the other way around. As a theoretical concept, the Input Hypothesis was widely discussed when it was proposed by Krashen in 1981 and during the following decades, but "as a theory, no longer figures in current thinking in SLA" (R. Ellis & Shintani, 2014, p. 176). VanPatten and Benati (2015) summarize in a similar statement, yet acknowledge that "*some* version of the Input Hypothesis" is found in all "major linguistic and psycholinguistic theories of SLA" (p. 129). Notwithstanding, the Input Hypothesis sparked a discussion that lead to an enhancement of theoretical hypotheses and teaching strategies by taking input into account as an indispensible factor for language acquisition.

Inherent in the critique on the Input Hypothesis is a notion that more emphasis should be placed on the learner's part in language acquisition, for if input remains unnoticed by the learner, it cannot be taken in. The Noticing Hypothesis (Schmidt, 1990) is predominantly concerned with the question whether there has to be a conscious process, or noticing, for language acquisition to take place. According to this hypothesis, mere exposure to certain linguistic features is not sufficient, but the learner needs to notice the linguistic forms in order to eventually acquire them. This infers a conscious process to take place when the learner is exposed to linguistic features, which is referred to as *noticing*.

Another focus in research has shifted to a perspective that takes conversational interaction into account, finding that negotiating meaning may contribute to the learning process, apart from speech modification and making speech comprehensible to learners of the language. This is the perspective of the Interaction Hypothesis, proposed by Long (1981) and continuously developed since (Gass, 1997, 2003, 2018; Gass & Varonis, 1994; Wesche, 1994). The Interaction Hypothesis also argues that comprehensible input is necessary, but a strong focus on interaction and the negotiation of meaning has been added as driving factors in acquiring a language. Comprehension problems between speakers lead to feedback, which in turn makes the learner understand as well as reconsider their language errors.

When looking at research on various forms of input and interaction, Gass (2003) summarizes that "in most cases non-native directed speech is grammatical albeit modified" (p. 249), implying that research on input and interaction theories rests on a native or near-native model speaker. As was shown in section 2.1, however, such a target language level in language

instruction may be found, but cannot be taken for granted in elementary schools in countries where English is a foreign language, such as the regular German elementary schools in the current study. Unlike the speakers Gass (2003) refers to, teachers' L2 input may not be grammatical. How interaction may be assisting in a learning situation that differs from the native speaker model and L2 learner has yet to be taken into consideration.

An additional aspect regarding the role of input in second language acquisition was brought forth for example by Swain (1985), who argued that input alone is not sufficient for language acquisition. She suggested that it takes comprehensible output as well, if a learner is to become a speaker of a foreign language, which is referred to as the Output Hypothesis. The background for the Output Hypothesis was that despite acquiring a foreign language in an immersion setting with a great amount of target language input, some of those children had acquired only little productive L2 language. The idea of comprehensible output is also integrated in interaction-based hypotheses in such a way that comprehensible output is considered to trigger the interlocutor to fine-tune their input. The interlocutor would receive the "signal" to "negotiate better input" (Skehan, 1998, p. 16).

One of the critical issues of the mentioned hypotheses is that they may obscure that comprehension makes one of more parts in language acquisition, but is not equal to language acquisition. Comprehension can take place based on non-linguistic features such as gestures alone. In her model of second language acquisition, Gass (2018) for example used the notion of *intake*, which refers to "the part of the language input that is internalized by the learner" (Gass & Selinker, 2008, p. 518). VanPatten's (1990) usage is similar: "Intake is thus defined as a subset of the input that the learner actually perceives and processes" (p. 287). According to this understanding, comprehension then is a necessary step to the acquisition process underlying intake, but does not refer to the same idea as intake.

Regarding the current thesis, incidental learning, frequency, and comprehensible input are relevant as theoretical hypotheses as in them, language performance may play a role. For incidental learning to occur, L2 teachers need to be able to provide a certain amount of target language that also shows a variety of frequent linguistic features and comprehensible input.

In order to understand which linguistic properties in the input are found to relate to the language learners' acquisition of the language, a perspective on the state of the art in first as well as second language acquisition is relevant. Therefore, the following sections examine what first and second language acquisition research has observed with regard to the relationship of linguistic features found in the input and children's language acquisition. Those insights form the theoretical background of the current study as they show if linguistic input and which properties of input relate to language

acquisition. Taking in a broader angle and including a brief summary of first language acquisition may put into perspective the role linguistic input has in language acquisition.

2.2.4 Input Effects in First Language Acquisition

Considering the linguistic properties of input in language acquisition research, various aspects have been examined in terms of first language development. In fact, research on how parental and caretakers' input affects young children's development goes back several decades (e.g. Hoff-Ginsberg, 1985; Lieven, 2010; Newport, Gleitman, & Gleitman, 1977; Snow, 1995). Experimental studies have added to the field of research on input effects and the development of particular structures (Brandt, Kidd, Lieven, & Tomasello, 2009; Lieven, 2010; Wijnen, Kempen, & Gillis, 2001). They found correlations between the prevalence of certain features in the caregivers' input and the children's acquisition of those features.

Some of the studies in first language acquisition research that have examined the relationship of linguistic features in the input and their acquisition in the children's language development, particularly analyzed syntactic or lexical features in the input with regard to the children's acquisition of those very features. Insights from those studies contribute to a general understanding of how input features can relate to language acquisition, yet bearing in mind that first and second language acquisition differ. Such a connection between first and second language acquisition has been drawn by Paradis et al. (2017) as well, who state that "since variation in the input determines individual differences in the rate of L1 acquisition of complex syntax in production, input factors could also be expected to drive individual differences in c[child]L2 acquisition of complex syntax."

As one of the earlier studies, Newport et al. (1977) found a significant relationship between the development of children's auxiliary fronted *yes/no* questions and their parents' use of them. Similarly, Hoff-Ginsberg (1985) found a positive correlation between parental *wh*-questions containing auxiliaries and their two-and-a-half-year-old children's acquisition of auxiliaries, which she traces back to the high frequency of auxiliaries in *wh*-questions.

A more recent study by Huttenlocher, Vasilyeva, Cymerman, and Levine (2002) looked at the relationship of syntactic features in the input of four-year-olds and the children's development of those very features. Huttenlocher et al. found that both the children's comprehension as well as their production of complex sentences was related to the proportion of complex sentences their parents produced.

Similar relations between linguistic features of caretakers' input and child language development were found in terms of lexical development and input. Pan et al. (2005) studied maternal lexical diversity and vocabulary growth of 1 to 3-year-olds. Their results found a positive correlation between the lexical diversity of the maternal input and the children's vocabulary growth.

By nature, however, the number of participants in natural parent-child interaction studies is limited, and "in many cases, reports of significant relations between parent speech and children's syntactic development have not been replicated in all studies" (Huttenlocher et al., 2002, p. 341). The same is true for other linguistic features in the input such as lexical diversity and how it relates to the children's vocabulary development. In total, there is support for suggesting that linguistic features in the parental input relate to the children's acquisition of those features, although research on their particularities, study replicability and comparability among the studies may be limited by the nature of the subjects and their individual environments.

Studies on the effects of child-directed speech have added to an understanding of how features in the input may relate to children's language acquisition. Infant- and child-directed speech differs from adult-to-adult speech in a number of characteristics, for example shorter utterances, shorter and less complex sentences, higher pitch, more exaggerated intonation, slower pace, and more repetitions (Pine, 1994; Soderstrom, 2007). The effects of infant and child-directed speech on children's language acquisition have been examined in a variety of studies on first language acquisition (e.g. Cameron-Faulkner, Lieven, & Tomasello, 2003; Lieven, 1994; Mani & Pätzold, 2016; Mintz, 2003; Pine, 1994; B. J. Richards, 1994; Rowe, 2008, 2012; Schreiner & Mani, 2017). Child-directed speech is suggested as being preferred by infants as it gives them a head start in segmenting the input (e.g. Cameron-Faulkner et al., 2003; Thiessen, Hill, & Saffran, 2005). Among the prosodic features such as intonation, pitch, and pace, pausing may be considered an aiding feature for children in segmenting language in the input.

In addition to the studies in first language acquisition research in which parental caretakers were the subjects, a few studies examined the linguistic input provided by institutional staff such as preschool or school teachers. Those studies add a valuable contribution to the understanding of input effects by taking into account an institutional setting. Huttenlocher et al. (2002), in their second study, analyzed preschool teachers' L1 language input and its effect on the children's syntactic development. They found that the teachers' speech did not significantly relate to the children's syntactic levels at the beginning of the school year, measured in the proportion of multi-clause sentences and the mean number of noun phrases per sentence, but was significantly related to growth over the school year as tested at a

later stage of the school year. Those results not only indicate a relationship between caretakers' input and the development of syntactical structures, but also that developmental stages differ in their suggestibility to input. According to those findings, children were able to benefit more from being provided with more complex structures later in their development. Huttenlocher et al. (2002) suggest the results "implicate the syntax of input providers as a factor that affects the extent of syntactic growth" (p. 370).

A different study carried out by Bowers & Vasilyeva (2011) with 104 preschoolers in the U.S. examined how features in the teacher's classroom input related to the children's lexical growth over the course of a year. The "number of word types was significantly and positively related to the PPVT [Peabody Picture Vocabulary Test] growth in English monolingual students but not in the ELL students" (Bowers & Vasilyeva, 2011, p. 233). Their findings suggest that the English as a first language speakers benefit from more diverse lexical input, whereas the L2 speakers in the group did not.

Studies on bilingual first language acquisition add valuable insights because the development of two languages can be observed and related to the input of each respective language. There is support for the assumption that home input of each language affects both lexical and morpho-syntactical acquisition rates of bilingual children (e.g. Paradis, Nicoladis, Crago, & Genesee, 2011). In addition, a large number of the studies examining input effects on bilingual language acquisition focused on the amount of exposure to each language. They found the amount of exposure positively affecting the rate of acquisition (Pearson, Fernandez, Lewedeg, & Oller, 1997; Unsworth, 2016b).

However, study results also suggest that the effect of the amount of exposure to each language can be moderated by low proficiency of the parental input (e.g. Chondrogianni & Marinis, 2011; Golberg, Paradis, & Crago, 2008). In these cases, the amount of language input did not make up for the parents' presumably low-level L2. The parents in the studies, who were themselves in a process of learning the L2, had spoken the L2 with their children at home too.

In sum, first language acquisition research shows support for input affecting children's language acquisition when morpho-syntactical and lexical features were examined in such a way that a highly diverse input can be expected to correlate with children's acquisition of those features. A large amount of input alone, however, may not suffice.

Two main theoretical considerations underpin the study at hand: first, the relevance target language performance has as a base for teaching, and second, how features in the linguistic performance interrelate with the acquisition of linguistic features in the subjects. As the present study participants taught and learned in an instructed foreign language setting, input effects specific to an instructional setting are reviewed in the following section.

2.2.5 Input Effects in Second Language Acquisition

Many of the aforementioned studies on first language acquisition looked at very young children up to the age of about three. Even though the psycholinguistic processing that takes place in a naturalistic language acquisition setting "presumably takes place in a classroom situation" (Gass & Selinker, 2008, p. 368) as well, instructed foreign language acquisition differs from early first language acquisition in several aspects: First, the learners have already passed age-related stages in their first language acquisition and have therefore gained crucial experience in processing language and in dealing with a variety of linguistic features in their first languages. Second, learners at a later age are more mature in all other areas of development – physically, cognitively, emotionally, socially, which can benefit the development in certain areas of language (Pica, 2005).

The most obvious difference yet between language acquisition in a natural environment and in an instructional setting may be the amount of input first or foreign language learners receive. Sharwood Smith (1994) states a calculation that a five-year-old child will have received about 9,000 hours of native language input. Whereas in naturalistic second language acquisition settings, where the L2 is widely spoken, the learners encounter the target language in their environment and are exposed to a large amount of second language input and a variety of speakers of the language, the instances when learners are exposed to the target language in an instructional foreign language setting may be largely restricted to the classroom. Instructed foreign language classroom learning entails a limited amount of time and a limited number of model target language speakers. Consequently, the amount of contact with the language drops considerably as well as possibly the quality of the language input, if there is only one language provider. Immersion programs, in particular those with a high amount of immersion, face a less restricted input situation since the language may be provided by multiple speakers on a variety of occasions over the course of the school days. A number of studies have found input intensity in terms of the amount of exposure to the target language to positively effect children's L2 development (e.g. Kersten, Schüle, & Steinlen, forthc. Lightbown, 2014; Maier et al., 2016; Muñoz, 2014; Rohde, 2010; Saito & Hanzawa, 2018; Steinlen et al., 2010; Unsworth, 2016a; Weitz, Pahl, Flyman Mattsson, Buyl, & Kalbe, 2010).

In addition to questions as to how input as linguistic data can be processed, research has been concerned with how particular modifications in the input could affect language acquisition. Model and learner language have shown to be mutually affective in a variety of settings: adult speakers may adapt their speaking level to the one of the children, which is referred to as child-directed speech (Clark, 2009), or native speakers may use

non-native directed speech to L2 speakers, referred to as 'foreigner talk' or non-native directed talk (Meisel, 1977; Gass, 1997; Gass & Selinker, 2008). Both child-directed speech and non-native directed talk are simplified variations in speaking, subconsciously or consciously intended to help learners understand and to keep communication and interaction ongoing. Modified, or adapted, target language can also be found in teacher-talk, which is a form specific to instructed second language learning.

2.2.5.1 Teacher-Talk

Similar to speech modifications observed in natural language settings, language may be modified in foreign language classroom as well, referred to as teacher-talk. Modified speech is characterized by a slower speech rate, more high-frequent words, simplified syntax such as shorter sentences, repetitions, and fewer clauses, discourse adjustments such as clearly connecting pronouns and their antecedents, and prosodic changes such as stress on content words (VanPatten & Benati, 2015, p. 100). Teacher-talk refers to this form of modification in a teacher's choice of linguistic forms and has commonly been referred to as the way teachers adjust their language to their students (R. Ellis & Shintani, 2014).

According to Wesche (1994), modified input can be helpful for lower proficiency learners of a foreign language. In her studies, elaborated teacher-talk has shown to increase learners' written and oral comprehension. In addition, R. Ellis and Shintani (2014) state that "teacher-talk may be ideal for lower proficiency learners but inadequate as a source of input for more advanced learners" (p. 189). Leow (1993), on the other hand, found no difference in comprehension if the input was simplified.

However, there is support for the assumption that if simplification in the input can help language acquisition, it is when a certain way of modification is applied. R. Ellis and Shintani (2014) refer to an earlier study which showed that in terms of vocabulary acquisition, it was the range as well as the length of the command in which the words were used that significantly correlated with the learners' vocabulary scores. A problem of teacher-talk, however, may be that it does not show a full range of grammatical forms and presents high-frequency words only (R. Ellis & Shintani, 2014, p. 189). Therefore, R. Ellis and Shintani (2014) not only mention the possible benefits of teacher-talk, but also note that there is "the danger that simplified input will deprive learners of exposure to the wide range of linguistic features needed for full development" (p. 188).

In consequence, even though teacher-talk may be referred to as linguistic simplification, it is in fact rather modification that could help students acquire or take in particular forms in the language. Modification may or may not include simplified forms.

Studies examining linguistic features in the teachers' language and children's development in the target language are still scarce. Milton (2009), for example, has focused on vocabulary acquisition and teacher-talk – among many other aspects of vocabulary acquisition – and points out that there is only little knowledge about teachers' language and its effect on vocabulary acquisition, despite the belief that teacher-talk greatly determines language acquisition (p. 212). He calls it "[t]he under-researched aspect of teacher oral input" (Milton, 2009, p. 213). In his study, Milton examined the words' frequencies with respect to how often they had been used by the teacher and states that "[r]epetition and recycling seems to have a beneficial effect on the likelihood that a word will be learned, but it is not, necessarily, an essential condition of learning" (2009, p. 211). He found that even those words that were rarely recycled were also learned in acknowledgeable numbers.

The idea that frequency in terms of word repetition and recycling, as was described in the frequency hypothesis in section 2.2.3, can positively affect the students' vocabulary acquisition is supported by a study by Donzelli (2007), who recorded numerous English classes in Italy. She found some indication that the frequency of particular words in the input predicted the learners' uptake of those. However, she found no evidence that words in the input from the lower frequency bands were more difficult to acquire than more frequent ones. Those findings suggest that it is not the nature of the measured general frequency that facilitates or retards acquiring words but their prevalence in the input.

Bowers and Vasilyeva's (2011) study investigated how the teachers' lexical diversity in the L1 input not only related to their L1 English speaking students' development in receptive vocabulary, as mentioned in section 2.2.3 above, but also to the L2 English learners in the group. They found that "the total number of words the teacher produced was significantly and positively related to the PPVT growth of ELLs, indicating that higher amounts of teacher speech correspond to higher rates of growth on PPVT" (Bowers & Vasilyeva, 2011, p. 232). On the other hand, the results showed that there was a significant negative relationship between the number of words per utterance and receptive vocabulary growth. Thus, a large amount of input was positively related, but only when the input was broken into shorter utterances.

Coming from a slightly different angle, Jones and Rowland (2017) set up a computational model that compared the effects of input quantity and lexical diversity in the input on children's vocabulary acquisition. Jones and Rowland (2017) recognize that individual differences of children's vocabulary development "are strongly predicted by environmental factors, particularly the quantity and quality of the linguistic input children receive" (p. 2). The reasoning for using a computational model was to avoid the

problem of confounding measures of input quantity and input quality. In their summary of the studies on amount and quality of input, Jones and Rowland (2017) found that usually, in particular diversity in the input and amount of input highly correlated. They argue that therefore, natural speech samples make it almost impossible to compare input quality and amount of input and analyze those independently of one another. In the computational model they applied, Jones and Rowland (2017) were able to independently manipulate the amount and the quality of the input. Their results suggest that while input quality was initially important, lexical diversity then outperformed input quantity as a predictor in the children's vocabulary development. Paradis (2011) came to similar findings when she investigated several external and internal effects on English second language learners in Canada. She found "rich L2 input" to be beneficial. However, what is considered as rich in that context might not be applicable to other settings: Richness in Paradis's (2011) study "included how much native-speaker input, as well as rich L2 input, children received" (p. 217). Paradis's study also exemplifies how much the term *rich input* depends on the context of the study, as the study design will determine the definition of *rich input*.

Whether explicitly stated so or not, the mentioned perspectives and hypotheses discussed in section 2.2.3 are based on native or near-native speaker models and the assumption that the model speakers of the target language are at a language level that allows them to modify and adjust the target language. This might not necessarily be the case as L2 English speakers may not only behave differently in the way they modify speech but may also not simplify at all if their language level is already low. One of Donzelli's (2007) conclusions was that "teachers are able to create a stimulating lexical environment – one that would encourage better chances for incidental acquisition to occur" (p. 122). Her study on how the teacher's in-class word use affected the acquisition of vocabulary of fourth-graders in Italy reveals this. The teacher participant, however, was also a native speaker of English building on 15 years of teaching experience in Italy – a combination Donzelli calls non-representative of the situation at primary level education in Italy. Such a combination would not be representative of the situation of primary school language teaching in Germany either, where the vast majority of English teachers are L2 speakers of English.

Milton (2009) adds another crucial idea why teacher language should be studied as relevant factor at a certain age of acquisition: "The relevance of the study of the oral language of teachers takes on a greater salience when it is considered how many children now learn a foreign language at an age when they are still learning to cope with reading and writing in their first language" (p. 212).

In short, teacher-talk is specific to instructed language learning. Whether teacher-talk differs fundamentally in non-native low proficient teachers on

the one hand, and non-native highly proficient and native speaker teachers of the language on the other hand, remains a subject to future research. The following section examines the relationship between teachers' language performance and second language acquisition in rather scarcely existing study findings.

2.2.5.2 Teachers' Language Performance and Second Language Acquisition

In second language acquisition research, input as language stimuli in general, and specifically in terms of the teachers' language performance, has not been under much scrutiny, neither as a research domain in its own right, nor as a factor with a possible effect on the development of a second language. A general significance of foreign language teachers' proficiency for second language acquisition has been proposed frequently, albeit often suggested rather than analyzed, and not necessarily based on any specific proficiency scores. Medgyes (1992), for example, states that "the ideal non-NEST [non native English speaker teacher] is the one who has achieved near-native proficiency in English" (p. 349). Notwithstanding, it may seem reasonably obvious that higher teacher L2 proficiency may be beneficial to learners' L2 language development. Yet few studies have analyzed why target language proficiency may be relevant to teaching and learning a foreign language.

The reason may be only to a lesser account its lacking potential as a considerable factor in language acquisition. After all, input is unequivocally considered a prerequisite of any form of language acquisition. One suggested reason why teachers' language and its impact on the students is rarely at the focus of research was that looking at teachers' language proficiency could be considered a "delicate issue", as Nikolov and Mihaljević Djigunović (2011, p. 107) call it. Teachers cannot be expected to readily volunteer in any study that specifically has in its focus their proficiency, as it is likely they will feel being put to the test.

Studies that have reported on teachers' second language input have looked at input in terms of didactic measures and interaction strategies, for example Weitz, Pahl, Flyman Mattsson, Buyl, and Kalbe (2010) and Weitz (2015) for preschool teachers, or the amount of first and foreign language choice used in the classroom (e.g. Inbar-Lourie, 2010). Rating classroom input on scales can provide valuable insights into the teachers' teaching strategies including their target language use. Observation schemes have been developed to capture a variety of classroom features (for a summary see Mackey & Gass, 2016). The Input Quality Observation Scheme (IQOS), for example, was used to investigate how features in the input relate to preschool children's target language development (Weitz, 2015; Weitz et al., 2010). The observation scheme also includes a rubric called "varied input"

(Weitz et al., 2010, p. 50). Judgments are dependent on the individual raters and as such, their own language skills and what they regard as varied or rich language. The limitation of rater subjectivity has also been acknowledged in the study (Weitz et al., 2010, p. 23).

In particular if the raters speak the same first language as the observed subjects and English as a foreign language as well, the rating is likely to render some bias. Saito and Shintani (2016), for example, showed how much raters' own language backgrounds affected their judgments of L2 English speakers' comprehensibility: The Canadian monolingual raters and the Singaporean English-speaking raters judged the presented speech samples not only differently but were also influenced by different phonological or lexical and grammatical features in the L2 speakers' samples.

Apart from the challenges of rater reliability in rater rubrics, "varied" in terms of linguistic properties is compelled to remain rather vague. Skehan's (1998) phrasing of "good quality input" (p. 17) is more interpretative yet as it is based on the observer's subjective idea of *good*.

Only a small number of studies have attempted to relate the teacher's language proficiency to the students' linguistic development. When studies took teachers' language proficiency into account, the results were commonly based on self-assessments, native speaker judgments, or (self-) reported English proficiency test results (Butler, 2004; Loder Buechel, 2015; Nel & Müller, 2010; Unsworth et al., 2015; Van Canh & Renandya, 2017).

Unsworth et al. (2015), for example, found teachers' proficiency, measured in native-speaker ratings, to be a predictor of the outcomes of the children's receptive vocabulary and grammar scores. Teachers' language proficiency rather than weekly exposure time was the best predictor of children's scores. Unsworth et al. (2015) compared the groups of students who had either (a) a non-native teacher, (b) a non-native plus a native teacher who taught jointly, or (c) a native speaker teacher. The children with a non-native speaker teacher at a level B of the Common European Framework of Reference [CEFR] scored the lowest on the tests. The group who had a non-native teacher at the same level co-teaching with a native speaker scored higher, as did the groups who were taught by a higher CEFR level teacher or a native speaker. Hence, Unsworth et al. (2015) conclude that native speaker input makes up for lower level non-native speaker input. They also suggest that input from a native speaker is not generally essential to progress in language development.

A study on the amount of input and teacher proficiency in L2 French was done by Graham et al. (2017) with UK school children in grades five, six, and seven. Teacher proficiency was based on the teachers' self-reported highest level of French qualification ranging from no formal French qualification, General Certificate of Secondary Education (CEFR level A), A level (CEFR level B2), university degree (CEFR level C2), and

native speaker. Graham et al. (2017) reported that, while being difficult to determine what the "optimal level of teacher language proficiency is" (p. 929), the teachers' French proficiency related to the students' outcomes on the tested grammatical and lexical features of French, as did amount of instruction time.

At this point in time, it remains inconclusive how teacher proficiency with its broad range of definitions and applications in the studies relates to second language acquisition and if factors such as amount of instruction may mediate an effect. Contrary to the limited amount of data on teacher language in second language acquisition, however, in studies on first language acquisition – monolingual as well as multilingual – databases such as the CHILDES project (MacWhinney & Snow, 1985) allow access to real-time recordings and transcripts of caretaker and child communication for analysis. What has been done for first language acquisition research, namely a linguistic account of the primary language providers, is virtually lacking in instructed second language acquisition research. In a recent study, Rankin and Unsworth (2016) state that "the need to take a more robust empirical approach to input is clear if we are to develop a deeper understanding of the nature of input effects" (p. 564). As their study is a reply to a claimed negligence of generative approaches to address input, they add: "both in terms of POS [Poverty of Stimulus] effects and also in terms of distributional properties of the input available to L2 learners" (Rankin & Unsworth 2016, p. 564). Poverty of Stimulus refers to what is considered the logical problem of language acquisition. Researchers of first and second language acquisition alike have been studying and discussing what is called the logical problem of language acquisition, which asks the question "how acquisition could work in principle – how a learner can correctly generalize from a finite sample of sentences in context to the infinite set of sentences that define the language from which the sample was drawn" (Pinker, 2004, p. 949). Research in those fields has been occupied with the psycholinguistic process of language acquisition and development. By nature, the field of second language acquisition is just as concerned with the implications any insights into language processing could have with respect to teaching. [13]

Even though the studies accounting for teacher L2 proficiency are limited in number, there is some indication that – in line with common belief – proficiency may be beneficial for learners' L2 development. Yet more research is needed. An additional link between teacher L2 proficiency and learners' L2 development can be found in how teachers' language proficiency relates to their teaching strategies. The following section explores this relationship.

13 For a discussion on UG in L2 acquisition compared to L1, see e.g. Meisel (2011).

2.2.5.3 Teachers' Language Performance and Teaching Strategies

Within the practical professional field of teaching English as a foreign language, the role of performance is considered in different terms: One debate revolves around the question whether non-native speakers or native speakers are more, or less apt to be teachers of English, which may become relevant as to whom to employ as teachers. A few studies have focused on non-native teachers, such as Árva and Medgyes (2000), Medgyes (2001), and Llurda (2004, 2006). Their main interest, however, was not language proficiency but rather what the differences and commonalities were between native and non-native teachers' teaching behavior in the classroom. Medgyes (1992) proposes in his native/non-native teacher comparisons: "[I]n a purely non-native context, it looks as though 'The more proficient in English, the more efficient in the classroom' is a valid statement" (p. 347). Such comparisons have not been picked up much in recent research, arguably because regardless of how nativeness versus non-nativeness determines teaching strategies, reality has made such a dichotomy if not redundant, so at least without consequences. The imperturbable reality is that the number of English teachers whose first language is not English has been increasing parallel to the rising demand of English instruction around the globe (see Jenkins, 2015). By now, the majority of L2 language learners is being taught by teachers and instructors who are foreign or second language speakers of English as well.

In her review on teacher's oral target language proficiency and teaching, Chambless (2012) states that "there is no research that establishes a direct connection between teachers' TL proficiency and effective teaching" (p. s154). She continues that one of the main impediments is the lack of a base definition of effective teaching. Effective teaching can relate to teacher qualifications, instructional practices, the product as in students' learning outcomes, or a hybrid of several elements.

While teaching effectiveness and subject knowledge is an issue in any form of teaching, language proficiency is specific to language teaching. However, there is no generic understanding of language proficiency on which research may be based, as the contexts of the studies will determine what definitions of proficiency to follow.

Numerous descriptors have been put in place in proficiency frameworks. The Common European Framework of Reference for Languages [CEFR], for example, is the framework most commonly referred to in Europe, whose purpose it was to find common and comparable grounds to describe an individual's language skills in any European language. The CEFR recognizes six main levels: A1, A2, B1, B2, C1 and C2. However, the CEFR is not geared towards claiming standards for foreign language teachers but rather offers descriptive levels for all the European

languages alike. The scales are user-oriented and based on can-do statements describing how the individual may be able to communicate. Guidelines for teachers at German schools sometimes refer to the CEFR levels, when they state that a C1 level in the target language is expected for teachers of the language (Kultusmisterkonferenz KMK, 2013). However, English teachers at elementary schools in the state of Lower-Saxony are not required to prove a particular target language level.

Another framework was developed by the American Council on the Teaching of Foreign Languages [ACTFL] in the US, where becoming foreign language teachers are required to demonstrate their speaking proficiency in an interview. Their proficiency is assessed according to the proficiency guidelines set up by the ACTFL (*ACTFL Proficiency guidelines*, 2012). The ACTFL guidelines distinguish ten different levels of proficiency: "novice", "intermediate", "advanced", and "superior", of which the first three are additionally subdivided into "low", "mid", and "high". The Council sets the level requirement for foreign language teachers of English as a foreign language and other languages to "low advanced". The background of proposing proficiency guidelines is based on the assumption that target language proficiency interrelates with those teaching strategies which are suggested to be beneficial to foreign language learning.

How language proficiency can support teachers to adhere to beneficial teaching strategies has rarely been studied. A summary of those second language teaching techniques that are considered to aid second language acquisition follows to then be able to show how linguistic performance may interrelate with beneficial teaching strategies. They are compiled as part of a classroom observation scheme in Kersten (2019, p. 50f.) as follows:

(1) Cognitively stimulating activities within learners' realm of experience: meaningful content goals / language use / conversational goals, require specific linguistic items, active problem-solving, prior world knowledge, constant learner activation, authentic materials / realia, texts, genuine interactions, opportunities for genuine output, differentiation, demonstration, introduction of goals
(2) Verbal input (phonological, lexical, morpho-syntactic): high language proficiency, exclusive L2 use, high amount of L2 input, adapted speech rate, intonation, pauses, recurring phrases / formulas, verbal routines / rituals, repetitions, lexically and morpho-syntactically varied L2 input, comprehension checks, adaptation to different learners
(3) Non-verbal input (internal / external): body language, visual illustrations, hands-on materials, written labels / phrases, displays
(4) Promoting learners' output: waiting time, encouragement to use the L2, questions with open answers, allowing learners' L1 use, allowing non-verbal ways of expression, key vocabulary / phrases for learners' output

(5) Reaction to learners‘ output: focus on linguistic form of learners’ output within a meaningful context, appreciation, correction of content and language errors (explicit corrections, recasts (i.e. correct repetition of learners’ incorrect utterance) and prompts which lead to self-corrections of the learners). (Kersten, 2019, p. 50f.)

While discussing the specific nature of each teaching strategy that is assumed to be beneficial in detail is beyond the scope of the current study, the relationships between language proficiency and some of the teaching strategies are relevant for an understanding of how proficiency and teaching may relate. As an exemplary study in the field, H. Richards et al. (2012) investigated how the teachers’ target language proficiency related to their teaching strategies. The teachers taught a range of foreign languages in New Zealand. Their proficiency was rated according to their perceived proficiency level, the length of time they had studied the target language, and whether they had taken an international exam. H. Richards et al. (2012) analyzed the classroom behavior, observing the provision of appropriate target language, appropriate corrective feedback, use of the target language for classroom management, provision of meaningful explanations of vocabulary and grammar, provision of “rich language input” (p. 241) and ability to improvise.

H. Richards et al.’s (2012) results indicated several relationships between the teachers’ target language proficiency and their teaching strategies: The lower-proficiency-level teachers demonstrated corrective feedback, but mainly only on those grammatical features they were teaching. Those teachers were less able to provide feedback in other areas. For example, they were not able to provide correct pronunciation of new words the students asked about. The more proficient teachers, on the other hand, also more consistently explained vocabulary and grammar meaningfully than the less proficient teachers. In addition, the more proficient teachers used a wider range of vocabulary and linguistic structures and responded spontaneously to students’ questions. In terms of classroom management, the more proficient teachers used more varied phrases instead of fixed phrases for classroom management strategies, such as assigning the students into their workgroups, disciplining, and praising. While H. Richards et al. (2012) point out that the lower proficient teachers were able to teach the language to some extent according to their key teaching strategies, H. Richards et al. conclude that “teachers need to have an advanced level of TL proficiency so they can also provide meaningful explanations, rich language input for learners and respond spontaneously and knowledgeably to their learners’ questions on language and culture” (p. 244).

A relationship has also been stated between teachers’ proficiency in the target language and confidence in speaking the language (e.g. Butler, 2004).

In terms of self-efficacy, teachers' perceived efficacy was found to positively correlate with their self-reported English proficiency (e.g. J. C. Richards, 2017). The same was reported in Eslami and Fatahi (2008), whose results showed a positive correlation between the teachers' perceived efficacy and their self-reported English proficiency, as well as in Chacón (2005) and Ghasemboland and Hashim (2013). However, since both self-efficacy and English proficiency are based on self-reporting, there is some danger of the variables measuring the same underlying construct.

From the small array of studies incorporating teacher language proficiency, teachers with lower-level language proficiency can reasonably be expected to be less affective in their teaching strategies than teachers with a higher-level proficiency. Cullen (2002) explains the effects of limited proficiency on the teaching as follows:

> A teacher with a poor or hesitant command of spoken English will have difficulty with essential classroom teaching procedures such as giving instructions, asking questions on text, explaining the meaning of a word or replying to a student's question or remark. (p. 220)

The reverse, however, may not follow: A high level of proficiency may not automatically lead to more excessive use of beneficial teaching strategies. A particular language proficiency threshold, at which language proficiency can be beneficial in combination with supporting conditions and teaching strategies, has not been found.

Even though considered promising research, Chambless (2012) comments that it would take a "coordinated effort by multiple research teams using both qualitative and quantitative research methodologies" (p. s154) to understand how teacher proficiency, teaching effectiveness as observed in supportive teaching strategies, and students' learning relate. High proficiency may result in flexibility in language use and the ability to tune into the students' needs more sensitively.

2.2.6 Assessing Receptive Grammar and Vocabulary in Early Second Language Acquisition

Having discussed features of language performance, theoretical models and hypotheses about second language acquisition as well as the relationship between those issues and children's language acquisition, attention needs to be brought to what is measured on the part of children's L2 development that can serve as a base for establishing such relationships. In the field of second language acquisition, much less research has been carried out with

children than with adults (Oliver & Azkarai, 2017). Assessing children's early L2 development may demand different test formats from the ones administrable with adults. Research that incorporates findings on children's development in a second language is primarily based on standardized test results. The following is an account of how children's early L2 acquisition is assessed in differing environments and what studies have found on students' development of L2 receptive grammar and vocabulary.

This section outlines research examining children's early receptive L2 vocabulary and grammar acquisition in various language acquisition contexts with a focus on studies that have used similar tests to the ones in the current study. As tests on language production at an early stage of language development, in particular of young children, are likely to result in floor effects, tests on their L2 language development predominantly measure receptive areas of language (Unsworth et al., 2015). Receptive language skills are frequently reported to outperform the productive skills in language acquisition (Webb, 2008).

Vocabulary as well as grammar development can be considered indicative of general language acquisition. While it is still a matter of debate how exactly the acquisition of grammar and lexis trigger one another (for more, see Kawaguchi, 2013), it is undisputed that "while it is not yet clear that the vocabulary knowledge is driving the other aspects of language development, vocabulary certainly appears to develop in size and depth alongside every other aspect of language" (Milton, 2013, p. 75). The lexicon's role is deemed crucial in L2 development as it is argued to set off other aspects of language development such as grammar development (e.g. Ellis, 1997). Aiming at answering to what extent vocabulary relates to other aspects of language performance and functions as a predictor of the four skills of reading, writing, listening, and speaking, Milton's study (2013) showed that vocabulary "explains up to 50% of the variance in performance in scores gained from tests of the four skills" (p. 71).

As examples of tests on receptive language, the Peabody Picture Vocabulary Test (PPVT) (Dunn & Dunn, 2007a, 1959; Dunn, Dunn, Bulheller, & Häcker, 1965) as well as the Test of Receptive Grammar (TROG) (Bishop, 1989, 2003), have been widely in use in first language acquisition research including bilingual language acquisition (e.g. Bialystok, Luk, Peets, & Yang, 2009; Carroll, 2017; Oakhill & Cain, 2012; Smithson, Paradis, & Nicoladis, 2014). They have also been administered in studies on specific conditions in which assessment of productive skills is inadequate, such as in stuttering (e.g. Ryan, 1992).

Early instructed second language acquisition studies that used receptive grammar and vocabulary tests are less numerous. The studies incorporating receptive tests in early instructed school or pre-school second language acquisition have examined the fields of receptive grammar and vocabulary

acquisition either as the focus of research in itself, or as indicators of the level of language development (e.g. Buyl & Housen, 2015; Couve de Murville et al., 2016; Hopp et al., 2018; Horváth & Nikolov, 2007; Jaekel et al., 2017; Maier et al., 2016; Rohde, 2010; Schelletter & Ramsey, 2010; Steinlen et al., 2010; Steinlen & Piske, 2016; Steinlen & Rogotzki, 2008; Unsworth et al., 2015).

A common procedure to test receptive grammar and vocabulary applied across a variety of studies is the use of picture-pointing tasks. They can accommodate for one of the main difficulties in assessing children's foreign language at the elementary level – the fact that there is little foreign language production to gather, as for example Unsworth et al. (2015) note, and the possibility to still collect data that reflect foreign language development. For a standardized assessment of receptive vocabulary, the PPVT and its British adaptation, the British Picture Vocabulary Scale (BPVS) (Dunn et al., 2009; Dunn, Dunn, & Whetton, 1982; Dunn, Dunn, Whetton, & Burley, 1997), each in various editions, have been in use to assess receptive vocabulary in a variety of contexts including instructed second language acquisition. For receptive grammar the TROG and more recently also the European ELIAS Grammar Test (Kersten et al., 2010; Kersten, Piske, et al., n.d.) have been in use. The test results of those four tests are calculated in numeric scores based on preset formulas. Thus, they are highly reliable and to a great extent comparable across studies. Those two tests that were used in the present study are described in detail in section 3.3.1.2.

Across existing studies, however, the contexts in which the respective children acquired the second language vary considerably. Age, school type, and out-of-school exposure to the target language are some of the diverging factors that need to be mentioned if comparisons between the studies are to be drawn to make general statements about children's' L2 receptive grammar and vocabulary development.

Depending on the country they live in, early second language acquisition study participants of the same age may still be attending pre-school or have already entered primary school. The line between pre-school and elementary school is drawn depending on the respective country's educational regulations. As schooling ages differ between countries, children start English instruction as early as four, for example in the Netherlands (e.g. Unsworth et al., 2015), whereas in other studies, four-year-olds still attended preschools, for example in Germany and the UK (e.g. Rohde, 2010; Schelletter & Ramsey, 2010; Steinlen et al., 2010; Weitz et al., 2010). The four-year-old preschool children were usually not exposed to formal language instruction but rather experienced English as communicative means in every-day situations. Thus, four-year-old study subjects may be observed in pre-school settings, which differ from school setting of children at elementary schools. German pre-schools, for example, typically do not include a formal classroom setting in which subjects are taught by

a teacher, whereas a school setting is predominantly based on classroom instruction. This difference may be vital with respect to the second language development of children and needs to be considered if comparisons between groups of children are based on age. Children at the same age may have been exposed to the second language in different instructional and non-instructional settings, which may affect any measured language outcomes of the children's second language.

Similarly, children have been studied in different types of schools. The elementary school children who were studied at ages six and up were often part of immersion programs (e.g. Buyl & Housen, 2015; Couve de Murville et al., 2016; Maier et al., 2016). Immersion and bilingual school teachers may be more compliant to take part in research, as both types of school require a strong commitment to teaching in the target language. There may also be a stronger self-selection bias of the teachers, because they may be less self-conscious about speaking the target language than regular school teachers. On the other hand, regular school English teachers working in an environment where the target language is not predominantly spoken, may be less exposed to the target language, speak it on fewer occasions and less routinely than bilingual or immersion school teachers, unless they actively engage with English speakers or function in the target language on out-of-school occasions.

Children's out-of-school exposure to the target language may also differ between study participants, which is another contextual factor influencing second language acquisition. Language acquisition in a naturalistic second language environment differs from instructed foreign language acquisition: Exposure to the target language will be higher and more varied in an environment where the language is spoken, whereas exposure in an instructed foreign language context can be limited to one teacher speaker for a limited amount of time per week. In addition, countries differ greatly in how much English is integrated into everyday life, in particular in the media. To take the European Union as an example, English language TV programs are dubbed into the official language in some countries, for example France, Spain, and Germany, and broadcast in English with subtitles in others, such as the Netherlands and the Scandinavian countries (Media Consulting Group, 2008; Pedersen, 2011). Out-of-school exposure to English will usually be higher with English language media and – due to different native speakers in a variety of interactions on the respective programs – more varied for children growing up in one of the latter countries than those living in an environment in which English is typically limited to an instructed foreign language classroom. Enever (2011), for example, found that out-of-class exposure to French or Spanish in the UK is substantially limited, as opposed to English in Sweden, where exposure to English in comparably high. She also found undubbed TV programs in

English in the Netherlands and Croatia to largely contribute to the amount of out-of-class exposure to English, which positively affected the children's target language development. Similar findings indicating that specifically subtitled TV had a significant positive effect on children's L2 English development, are reported in several studies (e.g. Huang, Chang, Zhi, & Niu, 2018; Kuppens, 2010; Lindgren & Muñoz, 2013).

Studies have suggested that young learners are particularly prone to acquiring second languages implicitly, whereas older ones, who are cognitively more mature, can make use of explicit instructions more effectively (e.g. DeKeyser, 2000; Muñoz, 2008). It follows that a large amount of linguistic input, either in terms of hours of instruction, or immersive and outside-of-class contact with the target language and a variety of speakers, is necessary in order to serve the implicit learning of young children and may affect children's L2 outcomes. For example, Muñoz (2014) found that her variables *current informal contact* and *hours of immersion abroad* were stronger predictors of the students' language development than the hours of instruction, which "highlights the importance of contact with native speakers and exposure to input that is linguistically rich" (p. 476). Alcañiz and Muñoz (2011 in Enever, 2011) suggest that a higher amount of input exposure in school may level out low out-of-school exposure.

As studies using similar tests to the ones administered in Study 2 could contribute to the discussion of the present findings, some of the results are considered of those studies that used one of the receptive tests with children, the TROG or the ELIAS grammar test, or the BPVS or PPVT vocabulary test, bearing in mind that learning settings may have differed.

Children in bilingual settings have been reported to improve their receptive grammar and vocabulary significantly over time, the intensity of input being more predictive than time span of exposure (e.g. Schelletter & Ramsey, 2010; Steinlen & Rogotzki, 2008). Couve de Murville et al. (2016) also found that students attending bilingual schools improved their scores on the BPVS2 on average over the course of a school year. However, the individual groups differed – the group with the highest amount of English exposure improved the most among their age group.

Striking differences can be found with respect to individual test scores in all four receptive tests: In their preschool participants Steinlen et al. (2010, p. 94) found "a large degree of individual variation in the data of the ELIAS grammar test." Likewise, Unsworth et al. (2015) found "considerable individual variation in the EFL children's scores on the receptive vocabulary and grammar tests" (2015, p. 543), using TROG2 for grammar and the PPVT4 for vocabulary, as did Couve de Murville et al. (2016) for receptive vocabulary on the BPVS2. Twenty-four of the participants in the latter study decreased in their vocabulary scores.

Between-subject variation can be attributed to numerous individual factors in second language acquisition, for example working memory, phonological awareness, language aptitude, L1, motivation, or learning styles, to name just a few. A number of second language researchers have devoted a focus specifically on how and why individual learners react and progress differently in their second language development (e.g. Dewaele, 2009; Dörnyei, 2005; Paradis, 2011; Skehan, 1989, 1991). A study with elementary school children who were also administered the BVPS and ELIAS Grammar Test examined the relationships between socio-economic status and children's receptive L2 development at immersion and regular elementary schools (Trebits & Kersten, forthc.). They found that school type was the stronger predictor of the children's receptive grammar and vocabulary development. In addition, Trebits, Adler, and Kersten (forthc.), examined cognitive factors such as working memory, phonological short-term memory, nonverbal intelligence, and phonological awareness as well as socio-economic status as factors relating to the receptive grammar and vocabulary development of a group of children attending German regular and immersive schools.

Granting all the differences in the research aims of studies using receptive grammar and vocabulary tests with children, fairly consistent results can be found regardless of the research aims and the examined factors. The studies' results regarding children's development of L2 receptive vocabulary and grammar can be summarized in mainly three aspects that will be of relevance for the study of this research: (a) the mean scores increase over time, (b) individual students' scores may not increase over time, and (c) there is considerable inter-individual variance. A larger amount of input appears to be advantageous for second language acquisition, as in bilingual and immersion contexts as well as in naturalistic second language environment and contexts supplying out-of-class L2 exposure.

In sum, studies on early instructed second language acquisition of children are still few in number and not necessarily alike in terms of the language learning settings of preschool versus elementary school, or immersion and bilingual versus regular school foreign language instruction, or language environment. Cross-study comparability is therefore limited and needs to acknowledge the different settings in which children are exposed to the target language.

With respect to study designs, investigating second language acquisition in an instructed language acquisition context in a community with little target language exposure has strong merits as N. Ellis and Collins (2009) point out:

> Unlike in L2 situations, where learners' most significant exposure to the target language usually takes place outside the classroom, rendering it challenging

> to identify and measure, in foreign language situations the significant (and sometimes near exclusive) exposure may take place in the language classroom, facilitating observations of the interaction between input factors and acquisition profiles.[14] (p. 333)

The present study employed two receptive tests – the BPVS3 (Dunn et al., 2009) and the ELIAS Grammar Test II (Kersten, Piske, et al., n.d.). Both have been in use to indicate children's L2 development, are feasible to administer particularly at the emergent stage of language development, when there is still little second language production, are consistent in the analysis, and therefore most reliable in their comparability.

2.2.7 Summary and Discussion

The concept of input has shown to be manifold and highly dependent on the research field and underlying research aims. For the purpose of the present study, *input* is referred to in terms of its linguistic properties in language production.

In models of second language acquisition, one focus has been on how input transforms into output, as is shown for example in Gass's (1997, 2018) and Leow's (2015, 2019) models of second language acquisition. Another focus has been on how input in language acquisition is either simplified, as observed in child-directed speech and teacher-talk, or should be modified to assist language acquisition, and how input intertwines with speakers and their teaching strategies. Largely, considerations in input and teaching strategies, such as modifying language, can be regarded as pedagogical in nature, when they focus on a specific classroom language behavior. They aim to explain how linguistic input can be altered and how fostering specific learning environments may support students' second language development. Underlying the theoretical considerations and the promotion of particular teaching strategies, however, is the assumption that the language providers have the necessary linguistic means at their disposal to vary the input they provide. All theoretical considerations are based on a model speaker teacher who is a native or highly proficient speaker of the target language. It remains subject to future debate if the same considerations apply to lower-level target language teachers.

14 Note that what N. Ellis and Collins call L2 in this context refers to second language learning in a naturalistic second language environment only and does not refer to the broad use of the term SLA, which includes any form of foreign or second language acquisition.

In terms of the theoretical hypotheses that have been suggested, there is common ground as far as linguistic input is considered. The frequency hypothesis as well as the idea of incidental learning is based on the necessity of re-occurring linguistic forms in the input. In order for incidental language acquisition to take place, teachers need to be able to provide the necessary environment. A high frequency of the forms in the input in a variety of different linguistic contexts may benefit the intake of such features. In order to provide such an input, the model speaker needs to be able to access a great variety of linguistic features and to produce them at demand and spontaneously – a capability that cannot be taken for granted in all L2 English teachers and that has seldom been the object of investigation.

Contrary to a natural acquisition setting, in which language predominantly functions as a means to interact and communicate in everyday life, institutionally instructed second language learning typically faces a more restricted realm of language contact. Foreign language classrooms tend to abide by operational rules and objectives as defined by the respective curricula and according teaching material, for example the school board's curricula for the state of Lower-Saxony (Niedersächsisches Kultusministerium, 2006, 2018).

Despite research gaps in understanding the connection between language providers and second language acquisition, or any form of language acquisition for that matter, it is assumed that particular properties of language input relate to children's L2 language acquisition. Looking at the studies on the relationship of linguistic input and language acquisition, there has been ample support from first language acquisition research and some support from second language acquisition research for proposing that input in terms of its linguistic properties can affect language acquisition. Sizable amounts of the target language and lexically as well as structurally diverse language in the input are expected to be beneficial for second language acquisition. However, what amount of target language input qualifies as *sizable,* is relative to the language acquisition setting of the research, as is the range of lexical and syntactical features that defines diversity.

There seems to be an understanding of adjusting input according to learners' levels in natural as well as instructed language acquisition settings. Yet the role of linguistic input properties proves to be one of the factors in language acquisition not fully comprehended. Attempts to grasp and approaches to examine input as a factor in language development depend on the perspective of the field. In addition, little to nothing is known about how language modification differs if both speakers are L2 speakers of the language, which is the most common situation in school language instruction settings in those countries where English is spoken as a foreign language, such as regular public elementary schools in Germany.

With respect to the properties of input, there is evidence from research on teacher-talk that suggests that simplified input can be beneficial to language acquisition. Yet there is some indication that *simplified* does not equate to *less varied* or *less complex,* but in fact more varied and repetitive in a range of different linguistic contexts. Therefore, modified talk of a teacher who is able to make use of a variety of constructions does not equate to teacher-talk that is little varied in all areas of language and repeats the same forms in the same linguistic contexts. The distinction between *simplified* and *modified* in such a way has not been made in research. Yet it seems vital in particular with regard to second language instruction, as modified input and linguistically diverse input alike are promoted to be beneficial for instructed second language acquisition. Clearly, simplified language on the one hand and more diverse input on the other hand pose a contradiction, or two ends of a continuum, if *simplified* does not entail modification in a sense that includes variety in the language features and its linguistic context.

If input needs to be lexically and structurally diverse to foster second language acquisition, the target language providers need to be able to model such an input. They need to have a linguistic choice readily accessible to them. If the input providers are able to produce lexically and structurally diverse input, this could then be considered an ingredient of foreign language teaching beneficial to the children's language acquisition, albeit by no means the only factor.

Research on teacher-talk of native teachers of the target language or of teachers who are considered highly proficient in the L2, is necessarily limited in its application to the language performance of possibly lower level L2 speaker teachers, who do not have at their disposal a broad range of syntactical and lexical complexity and accuracy which they can deliver fluently when deemed appropriate. This may not only affect the linguistic input properties provided in the input, but also the teaching.

Whereas there is no guarantee that a certain proficiency in the target language results in a particular quality of teaching, as was shown repeatedly in dichotomous native/non-native speaker teacher comparisons, it has been suggested that lacking speaking skills in the target language will negatively affect the teaching methodologies in the target language. Thus, with respect to the teachers' L2, there is reason to assume that teachers' target language performance is a factor in the children's foreign language acquisition, based on the assumption that lexical and structural diversity as well as beneficial teaching may be supported by target language proficiency.

Considerable consistency can be found in the means of standardized tests of receptive grammar, namely the TROG (Bishop, 1989, 2003) and the ELIAS grammar test (Kersten et al., 2010; Kersten, Piske, et al., n.d.), and vocabulary, namely the PPVT (Dunn & Dunn, 2007a) and the BPVS (Dunn et al., 2009, 1997) – each in respective editions. The tests may be used to

examine each area of receptive vocabulary or grammar in their own rights, as the test results may inform on specifics in the development of vocabulary and grammar. Receptive grammar or vocabulary have also been used as indicators of a more general early second language development to examine effects of internal or external factors on second language acquisition.

To conclude, there is still much demand for research on how to operationalize language performance and how particular features in the target language of the language providers may interact with children's development of that language.

2.3 Desideratum

As was shown, the theoretical background to the current thesis involves a multitude of research fields in second language acquisition. The core themes of the discussions on the areas of language performance as well as second language acquisition and linguistic input are summarized in the following section.

A need for more research is prevalent regarding how linguistic performance can be measured. Measurements and definitions in the field of CAF are still much the subject of theoretical as well as methodological considerations in capturing linguistic performance. An abundance of measures and methodological research approaches to language performance impede the replicability and comparability of the studies. The present study expands on the dimensions of CAF as a framework for language production as well as engages in the methodological enhancement of the measures linked to each dimension. A statistical model is trialed to make the numerous CAF measures utilizable for further analyses.

No consensus has yet been found in research whether complexity, accuracy, and fluency inevitably *have to* trade off in foreign language speakers' language production. A considerable body of research has been occupied with the trade-off question. It can be considered an issue that is ultimately relevant for second language learning and teaching in at least two directions: For one, existing or non-existing trade-off effects shed light on how a second language can be retrieved and produced. For another – which is not independent of the first direction – insights into learners' second language production may influence teaching methodologies and curricular decisions.

Discussing linguistic input and its role in acquisition has revealed that first as well as second language acquisition research consider linguistic input the sine qua non for any language acquisition (Mackey & Gass, 2015, p. 181). A large body of research on the effects of linguistic input properties has been and is still being conducted in first language acquisition, particularly

bilingual first language acquisition. In second language acquisition research, language input as such and how it may function in second language acquisition has been subject to SLA models such as Gass's (1997, 2018) and Leow's (2015, 2019). As a genuine interest in second language acquisition research is also the translation into second language teaching and learning, theoretical hypotheses have been stated and are continuously being refined that include considerations on input processing as well as pedagogical considerations on enhancing input for learners. Those include the hypotheses of incidental learning (Schmidt, 1994a), input noticing (Schmidt, 1990), comprehensible input (Krashen, 1985), frequency (N. Ellis, 2002), and interaction (Long, 1981).

A research gap is prevalent in how L2 teachers' language performance relates to early foreign language acquisition. A neglected part in the theoretical considerations about classroom teaching strategies and second language acquisition is the global change that has resulted in the majority of the English L2 teaching body being second language speakers of English as well. Studies acknowledging teachers as foreign language speakers of the target language with varying degrees of proficiency are notably scarce, as are studies on L2 teachers' linguistic performance or proficiency and their possible impact on students' target language development. With respect to the CAF framework in particular, teachers' linguistic performance has not been studied in such a relationship. Considering children's second language development, there is great demand for further research regarding how second language acquisition develops in young children who are exposed to second language instruction by L2 English teachers at primary school level with a small amount of exposure to the target language.

The thesis at hand incorporates the CAF framework and its linguistic performance descriptors to analyze language production of English teachers, who were L2 speakers of English themselves as well. It examines the possible CAF relationships, and their impact on the development of children's early second language acquisition, as assessed in their L2 receptive grammar and vocabulary.

3 Empirical Study

The study examined elementary school L2 English teachers' language performance and children's second language acquisition. The research questions and hypotheses as well as the design of the empirical study are presented in the following section. No hypotheses are predicted referring to those research questions that were exploratory in nature and which cannot be based on already existing theoretical and empirical studies.

3.1 Research Questions and Design

The following research questions and hypotheses guided the empirical study:

(RQ1) How do the L2 English teachers perform in terms of complexity, fluency, and accuracy?

Teachers' L2 language has not been studied in a CAF framework. Therefore, no prediction is made.

(RQ2) How do the teachers rate their L2 English language proficiency and the modification of their language use in the classroom?

As the English language proficiency level of the subjects was not tested, the teachers' self-rating regarding their English proficiency was not predicted. Language use in the classroom has been discussed as being simplified in terms of vocabulary, sentence structure, and features of fluency, which is referred to as teacher talk (section 2.2.5). Therefore, the following was predicted: (H2) The teachers report modifying their English language use in the classroom.[15]

(RQ3) How do the students' receptive English grammar and vocabulary develop over their fourth year of elementary school?

No prediction can be made on the basis of existing studies with respect to the specific scores. Studies differ with respect to test instruments and a variety

15 For a better orientation throughout the thesis, the hypotheses are numbered according to the research question they relate to. Not every research question is followed by a hypothesis.

of features in the learning contexts, for example bilingual and immersion preschools or primary schools, different amounts of instruction time, and varying amounts of English exposure outside of the class (e.g. Couve de Murville et al., 2016; Rohde, 2010; Schelletter & Ramsey, 2010; Smithson et al., 2014; Steinlen et al., 2010; Unsworth et al., 2015). However, studies have reported a positive L2 development (e.g. Couve de Murville et al., 2016; Unsworth et al., 2015). For this reason and because the students in the present study had received regular weekly English instruction between the two test times, the following was predicted: (H3) The mean scores in grammar and vocabulary increase between the two test times.

(RQ4) How do the student groups differ per teacher in their receptive English vocabulary or grammar attainment and development?

The teachers were randomly selected L2 speakers of English, who had German as their L1, and were teaching English as a subject at regular public schools in Lower-Saxony. As mentioned in section 2.2.5.3, there are no specific language requirements for teachers teaching English at elementary schools in Lower-Saxony, Germany. Thus, no prediction is formed.

(RQ5) How can the CAF dimensions be transformed into a scale that can be used for further analyses?

Studies in the CAF framework have used individual measures of each CAF dimension, as was shown in section 2.1.4 on how complexity, accuracy, and fluency are measured. Applying a Principal Component Analysis represents a novel angle to operationalizing the CAF measure into composite scores. Therefore, no hypothesis is suggested.

(RQ6) How do complexity, accuracy, and fluency in the teacher's L2 performance relate to one another?

Studies on the relationships of the CAF dimensions are inconsistent in their findings. Studies focusing on the CAF scores within individuals, however, have found that complexity, accuracy, and fluency developed alongside with one another (e.g. Vercellotti, 2012, 2015). The present study therefore predicted the following: (H6) All three CAF dimensions correlate.

(RQ7) How does the teachers' L2 English performance, as measured in complexity, accuracy, and fluency, relate to their students' L2 receptive vocabulary and grammar development?

Measuring teacher L2 performance in the CAF framework has not yet been reported, nor have studies measured teacher L2 performance in terms of CAF. There are no comparable studies to this date that are based on similar statistical procedures in the analysis of the individual CAF dimensions. However, studies have suggested that the language performance of teachers or other language providers have a beneficial effect on children's language acquisition in many ways. Features in the linguistic input of caretakers were found to correlate with children's first language acquisition of those very features (e.g. Huttenlocher et al., 2002; Jones & Rowland, 2017; Lieven, 2010). A positive effect of teachers' L2 proficiency on children's development of L2 has been found as well (e.g. Graham et al., 2017; Unsworth et al., 2015). In addition, over-all target language proficiency was expected to affect teachers' confidence, spontaneity in using the language, and delivery of linguistically diverse input, all of which has been argued to be beneficial to the students' L2 development (e.g. Butler, 2004; Eslami & Fatahi, 2008; J. C. Richards, 2017). Further, over-all target language proficiency has been argued to facilitate using teaching strategies that may assist children in acquiring a second language (e.g. Chambless, 2012; H. M. Richards et al., 2012). In addition, a correlation between the CAF dimensions was proposed in hypothesis H6. Therefore, the following was predicted: (H7) There is a positive relationship between the teachers' CAF performance as well as each of the individual CAF dimensions and the students' receptive grammar and vocabulary development.

(RQ8) If there is a relationship between teachers' L2 performance and children's foreign language acquisition, is there an additional effect by the classroom L2 use as rated by the teachers?

Theoretical considerations on teacher-talk characteristics have suggested that modified language, for examples in terms of pauses and slower speech rate in the classroom, may benefit children's L2 development (e.g. Weitz et al., 2010; Wesche, 1994). Therefore, the following hypothesis was suggested: (H8) The teachers' adapted L2 use in the classroom moderates a possible CAF effect on the children's receptive grammar and vocabulary development.

The study follows a mixed methods approach and is based on qualitative and quantitative data. Mixed method approaches combine features of qualitative and quantitative research approaches (R. Johnson, Onwuegbuzie, & Turner, 2007). Methodological triangulation of methods "reduces observer or interviewer bias and enhances the validity and reliability (accuracy) of the information" (D. M. Johnson, 1992, p. 146) (*Figure 2*). A qualitative as well as quantitative approach was needed to access the particularities of the teachers' linguistic performance, while

quantitative testing was able to tap into the students' L2 development over their fourth year of elementary school. Such a combination of methods allowed two independent data sets, represented in Study 1 and Study 2.

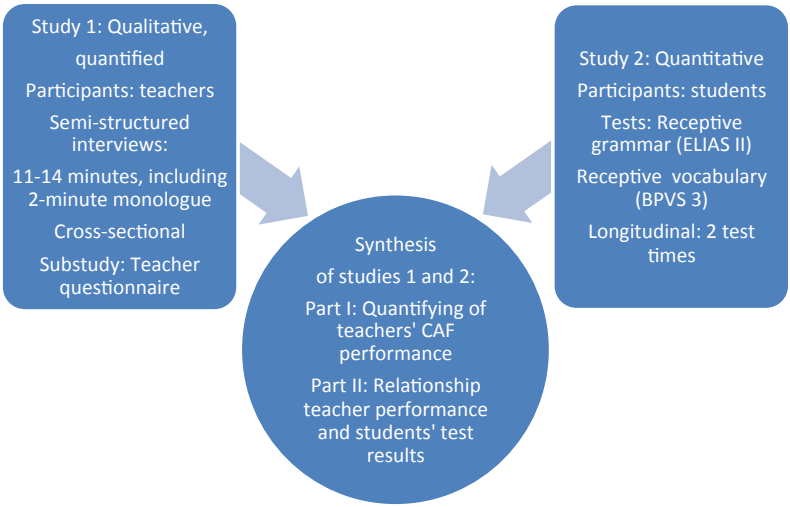

Figure 2 Triangulation of research methods

The studies are outlined in the following. Further details of each study concerning the design, participants, instruments, administration, reliability and validity as well as the data analyses are presented and discussed in the respective study sections of chapter 3.

Study 1 was based on semi-structured interviews with eleven English teachers. The results of Study 1 were made quantifiable in a coding process based on complexity, accuracy, and fluency measures. Quantification not only provides for detecting patterns of phenomena occurrence, but also helps verify and report patterns (Mackey & Gass, 2016, p. 234). Quantifying the data also prepared the interview data to be calculable in the synthesis section, which brings together both Studies 1 and 2 (*Figure 2*) in a numerical description. A questionnaire substudy was integrated to incorporate additional information on the teachers' self-rated classroom language. In addition to the indications that the linguistic properties in the teachers' linguistic performance are relevant to students' L2 acquisition, the modifications and simplifications in the English input in class could have a moderating effect on the students' foreign language development.

Study 2 incorporates the students' results of two standardized receptive tests on grammar and vocabulary longitudinally at two times during the fourth year at elementary school. The results of Study 2 were quantified by

the scores the students achieved on the respective vocabulary and grammar tests.

The synthesis Study 3 investigates if there was an independent relationship of teachers' L2 performance and students' second language acquisition. It was mandatory to measure the performance independently of a context exhibiting a number of confounding variables such as the classroom interactions, atmosphere, content, or methodological choices.

The impact of the teacher's performance as operationalized by CAF on their students' grammar and vocabulary development could be influenced by the in-class usage of the L2. Such a possible moderating effect was taken into account using the teacher questionnaire in a substudy of Study 1, which included self-ratings on the actual language use in the classroom.

Based on these findings, the current study can serve future research in developing and subsequently testing further hypotheses about teachers' L2 performance on the one hand and about its relationship with their students' second language acquisition on the other.

3.2 Study 1: Teachers' Language Performance

The first part of the empirical study, Study 1, is a linguistic analysis of the teachers' language performance. Eliciting real-time spoken language was considered to improve the ecological validity over laboratory settings. In addition, spoken mode was considered more valid than written language samples, as spoken language is the prevalent mode at the elementary school level (Niedersächsisches Kultusministerium, 2006, 2018). Data on the teachers' spoken English were elicited during one-on-one interviews with the researcher.

3.2.1 Data Elicitation Interviews

The prospective participants were approached via all means of communication: in person through the researcher, e-mails, or the school's principals, who then forwarded the requests to their teaching staff. All participants were a random choice of public elementary school teachers of English in the state of Lower Saxony, Germany.

3.2.1.1 Participants

The participants were eleven public elementary school English teachers. Lower-Saxony elementary schools cover grades 1 through 4, the usual age

of the children being about 6 to 10. All the interviewed teachers were L2 speakers of English with German as their first language.

The participating teachers taught English as an individual subject in an otherwise German curriculum. The English contact time was two 45-minute lessons per week and class. The teachers had varied teaching loads and taught other subjects as well. Table 1 shows the teaching experience of each teacher, whether they held a degree in English Studies, and if they had lived in an English-speaking country.

Table 1 *Teachers' English experience*

Teacher	Years of teaching experience	Degree in English	Stay abroad
1	2-4	yes	yes
2	2-4	yes	yes
3	>10	no	no
4	n/a	yes	yes
5	5-7	yes	yes
6	>10	yes	no
7	8-10	yes	yes
8	8-10	yes	yes
9	>10	yes	yes
10	2-4	yes	no
11	2-4	no	no

Note. Stay abroad in English-speaking country for at least three months.

In terms of the years of teaching experience, the participants varied from two to four years of teaching to over ten. Most teachers held a degree in English. Two teachers did not hold a degree in English. Four of the participants had not spent time abroad in an English-speaking country while seven had. Any personal identifiers of the participants needed to be discarded so as to protect the confidentiality of the participants' identities. This is particularly crucial when dealing with small sample sizes, as each study subject's identity may be easily linked to certain information (Information and Privacy Commissioner of Ontario, 2015). In the sample at hand, those identifiers included sex, age, and country name of their stay abroad and were therefore

not included in the table. The participants were between about 30 and 65 years old and 91% female. The ones who had stayed abroad had lived in an English-speaking country for a period between three months and a year. The countries included the USA, UK, Australia, and Ireland.

As expected, self-selectivity in a voluntary study on teaching English was considerable with respect to whether or not the participants held a degree in English. Most participants held a degree in English. Self-perceived language proficiency may have influenced the participants. In fact, one of the teachers who had been approached explained not wanting to take part in the study, because this teacher did not hold a degree in English and did not feel comfortable enough speaking English. In all other cases, however, no reason was given when teachers did not agree to take part. Therefore, it remains uncertain what exactly kept those who did not take part.

In sum, the participant sample showed a mix with respect to teachers with or without a degree in English, a broad age range, a broad seniority range from a few years of teaching experience to shortly before retirement, and experience in an English-speaking country. The sample included a highly unequal number of male and female teachers – a disproportion that reflects the higher number of female teachers at the elementary level in Germany. The sample can therefore be considered showcasing a fairly representative variety of English teachers at German elementary schools.

3.2.1.2 Interview Format

Because recruiting teachers to participate in a study conducted in English was expected to be a serious challenge, a one-shot interview design was chosen so as to keep the threshold low for the teachers to participate. The interview design was chosen with respect to its qualities to foster natural speaking behavior in a non-testing atmosphere. A conversational format was expected to help the participants provide extended speaking (Mackey & Gass, 2005, p. 174).

Semi-structured interviews were carried out in order to capture the individual teachers' speaking repertoire as extensively as possible and feasible. As opposed to structured interviews, semi-structured interviews are less rigid and allow the interviewer to move in participant-induced directions. They allow the interviewees to answer "on their own terms" (Edwards & Holland, 2013, p. 29), while the interview structure remains comparable across the interviews. Unstructured interviews on the other hand are the most similar to natural conversations but may be less similar in the language they elicit and thus render the sample data less comparable.

Several criteria further determined the format of the interview: time allowance, language choice, and topic. The allocated amount of time needed to suffice to let the participants become acquainted with the researcher and

the interview situation as well as to allow the participants to get into the mode of speaking English. Since the participants lived in a monolingual non-English environment and spoke English as a foreign language, setting the mode to English was considered a crucial criterion in eliciting as representative a data as possible, given the conditions. At the same time, the interviews had to be short enough to allow for efficiency and respect time restrictions on the interviewee's and interviewer's part.

In addition to the general outline of the interview format, the allocated time for the entire interview needed to allow a section of monologic speech that reflected the participant's ability to speak continuously for a certain stretch. A section of uninterrupted speech was needed to best apply the CAF measures, in particular fluency measures. Additionally, monologues have been reported to provide for more syntactically complex utterances than dialogues, for example in Michel et al. (2007). Yuan and Ellis (2003) add that monologues "afford a basis for deriving measures of learner performance that are not influenced by interactional variables" (p. 9), unlike dialogic narratives. Lastly and similarly importantly for the current study, "[t]he pragmatics of dialogues may interfere with the study of relationships among CAF" (Vercellotti, 2015, p. 19).[16]

The choice of language was English only. Since all of the participants were speakers of English as a foreign language and lived in a dominantly German-speaking environment, shifts to German on the interviewees' parts were likely to happen, in which case the data would not have been valid for the present study. Therefore, all conversations with the individual teachers were conducted in English.

The topic was intended to evoke some personally motivated language performance. Personal motivation was assumed not only to trigger willingness to speak, which is crucial in collecting spoken language data, but also to reveal a large part of the participant's language repertoire. Different topics may trigger different linguistic forms (N. de Jong & Perfetti, 2011). By choosing a school-related topic, differences in the participants' familiarity with the topic, which could affect their performance as well, were expected to be largely eliminated as all participants were teachers themselves. Different tasks can also affect the performances in terms of the CAF dimensions, as was the case for example in Foster and Tavakoli's (2009) study, in which differing storyline comlexity in the recorded narratives may have affected the CAF scores. The task format in the interview as well as the actual topic was the same across all participants to ensure comparability and to reduce possible task effects.

16 For more on dialogic versus monologic performance see Ferrari (2012) and Tavakoli (2016), for example.

A similar interview format is used in the International English Language Testing System [IELTS] (2007) speaking part to elicit spoken language. More freedom was applied to the current interviews, as they were not intended to test the participants. On the contrary, the setting was meant to foster a relaxed, comfortable atmosphere that acknowledged an as equal as possible relationship between the interviewer and the interviewee to empower the participants in their thoughts and speaking intentions.

The teachers were interviewed individually for 11 to 13 minutes, each interview consisting of three parts. Part one was a four to five-minute introductory dialogue between the interviewer and the interviewee about personal questions, such as where the interviewee grew up and how living there might have changed over time.

Part two was a long-turn monologue of about two minutes on a given subject presented by a stimulus (*Figure 3*). The teachers were given topic cards and a minute to prepare before they talked uninterruptedly for about two minutes. Approximately the same speaking length of two minutes has been used in a number of studies for the analysis of spoken language (N. H. de Jong et al., 2013; N. de Jong & Perfetti, 2011; R. Ellis & Barkhuizen, 2005; Vercellotti, 2015, 2019), while less speaking time of one minute (Tavakoli, 2016, 2018) to shorter twenty-second extracts of two-minute recordings (Pinget et al., 2014) have been used as well.

The topic was adapted from IELTS free practice material (IELTS, 2007, p. 12).

Describe a teacher who has greatly influenced you in your education.

You should say:

- *where you met them*
- *what subject they taught*
- *what was special about them*
- *and explain why this person influenced you so much.*

You will have to talk about the topic for 1 to 2 minutes. You have one minute to think about what you are going to say. You can make some notes to help you if you wish.

Figure 3 Topic card
(Adapted from IELTS, 2007, p. 12)

Asking the participants to reminisce about one of their own former teachers was expected to be relevant to the participants and to evoke a personal motivation to speak. It was also expected to be sufficiently interesting to prevent the participants from intentionally focusing on any one particular form of their language performance, which could then have impacted the

results. The halo effect – the participant's potential indirect adjustment to perceived expectations of the interviewer – is one of the caveats in data elicitation through interviews (see Mackey & Gass, 2005, p. 174). A familiar, personal topic and a participant-centered interview technique were chosen to limit the risk of a halo effect.

Another rationale for choosing a narrative format and this particular school-related topic was to motivate the participants to speak about background and foreground events as this can trigger sentence subordination, which is a determiner of syntactic complexity (see Foster & Tavakoli 2009). Narratives can involve "syntactically packaged constructions" (Berman & Slobin, 1994, p. 14), in which phases in the narrative are hierarchically sorted through subordination. Berman and Slobin (1994) argue that the reason why children acquire a complex system instead of maintaining simple sequences of clauses in their narratives is because children develop the capacity to construct a hierarchical order of events and apply subordinations and interrelationships between the events. Instead of picture stories as used by Berman and Slobin (1994) with children, the interview format in the current study involved narrative and episodic interviewing techniques to allow for person-centeredness (Bates, 2004). In the current study, *narrative* therefore refers to the personal narrative of the study participants.

Part three of the interview was a four to five-minute dialogue with the interviewer on the topic. In this part there was a shift to a more theoretical than personal base regarding school and education. The questions followed up on more complex issues such as how the interviewees regarded teachers' education, thereby offering the opportunity to include a less personally motivated speaking style, but a more argumentative language to support particular viewpoints.

Because of its monologic nature, the middle part of the interview was feasible to be used in the CAF analysis. In the set-up of an interview, the middle part is also the one that should provide the key questions, as the participants are beyond the nervousness of the beginning and not yet as tired as at the end (Mackey & Gass, 2005).

3.2.1.3 Interview Administration

All subjects were asked individually to take part in an interview and to give their written consent. They were informed that there would be no preparation for them, that the interview would take about 10 to 12 minutes and would be audio-recorded. A semi-informed interview format was used to elicit reliable data from the participants. The content details of the interview were not revealed beforehand – a common procedure in second language research, when giving away details of the aim of the study may influence the participants' replies and therefore result in unrepresentative

sample data (Mackey & Gass, 2016, p. 35f.). The participants were debriefed immediately after the interviews.

The interviewer met with each teacher individually in a quiet room of their choice, except one teacher. That participant was offered to do the interview via phone because the participant preferred not having to physically meet at a particular place at a given time. This interview was recorded via Skype internet voice and video call. By connecting Skype to the software application Audacity ("Audacity (R): Free Audio Editor and Recorder [Computer program]," 2015), that interview was recorded as well. The other ten interviews were recorded on a Zoom H2N Handy Recorder that was placed on the side of the table at which the interviewer and the respective participant sat face-to-face.

Before starting the interview and the recording, the subjects were informed about the length of the three parts of the interviews. They were also told that they would be given a task in part two.

Throughout the interview much attention was given to creating a comfortable, friendly and relaxed atmosphere. Anxiety needed to be kept at bay, in particular as some of the subjects either seemed to be or had in fact expressed being somewhat uneasy about speaking English with an adult native speaker. It is noteworthy that a certain amount of self-consciousness was observed in some of the participants even though the individual participation in the study was voluntary. This suggests that at least some participant variance with respect to the language performance was to be expected.

All of the eleven interviews were administered according to the preset format and set-up. Possibly existing self-consciousness at the start of the interview was noticeably diminishing already during the first part of the interviews, in which the participants readily spoke, seemed relaxed and genuinely interested to talk about their personal backgrounds. A large amount of interviewee speaking time and fewer cues from the interviewer demonstrate this (see transcripts Appendix A, Appendix I, Appendix J, Appendix K). The interviewer encouraged the participants by keeping eye contact, nodding and back-channeling *mhm* or *uh huh*. In all cases, the participants spoke for a considerable amount of time already in the first part of the interview, which lasted about four to five minutes. After the first part of the interview, the participants looked at the task card and took notes. They were asked to start speaking after about two minutes. While some put down the pencil immediately when asked to do so and started to speak, others were slightly reluctant to stop taking notes and were kindly asked several times to stop writing and were encouraged to speak.

The participants filled one to two minutes of monologic speech in part two of the interview. The average interview duration of the second part was 2.25 minutes (134.98s) with a standard deviation of 0.51 minutes

(31.71s). The interviewer did not interrupt during the monologic part. Backchanneling was done through nodding and keeping eye contact. In one case the interviewer intervened during the second part of the interview, when the participant intended to look for a dictionary. The participant was kindly asked to keep speaking without one and was encouraged to continue, which the participant did.

After the end of the recordings the participants continued speaking about the topic. Some expressed their surprise that the interview was over. Except for the participant in the telephone interview, all of the interviewees continued speaking to the interviewer about the topic for some time after the recorder had been turned off.

In sum, the interview format, the interview administration, and the topics were able to elicit a great amount of participant speaking so that in all the interviews the speaking time of the participants exceeded the interviewer's speaking time by far, as is sampled in the interview transcripts of four teachers (Appendix A, Appendix I, Appendix J, Appendix K). The language data can therefore be considered to a large extent authentic and representative of the linguistic skills of each participant. This was important in order to validly conduct the analyses and to then be able to make claims about the linguistic performance of the participants. Complete transcripts of four teachers, Teachers 1, 9, 10, and 11, were included as illustrations of the full-length interviews (Appendix A, Appendix I, Appendix J, Appendix K). The students of those four teachers' classes participated in Study 2.

3.2.2 Data Analysis Interviews

In contrast to test scores, which can be immediately used for analysis, natural data first need to be transformed into an analyzable format (Mackey & Gass, 2016, p. 112). The following sections illustrate the steps taken and the choices made on transcription conventions and data coding.

3.2.2.1 Transcriptions

The recorded interviews were transcribed with the following cautionary note in mind:

> Because it is impossible to document all features of social interaction, all transcripts should be considered *partial* representations, *selective* and *situated* in relationship to the goals of a particular study (Davidson, 2009; Lapadat & Lindsay, 1999). (Paulus, Lester, & Dempster, 2014, p. 95)

The interview transcriptions are verbatim transcriptions. Verbatim transcriptions include "all utterances made by all participants without changing non-standard language usage (e.g. 'he don't care about me') or dialect […] and without skipping over repetitions ('and-, and-'), false starts ('uh-, well, I mean') and backchannels ('mm-hmm')" (Paulus et al., 2014, p. 96). A particular English dialect did not apply to any of the present participants, but certain features of German were prevalent in their English, most notably German lexical items and sentence structures. The German words were transcribed in the original German form and translated into English in double parentheses.

Choosing a verbatim transcription guaranteed that all the required linguistic output was included to prepare the data for the interview analyses to be performed. The CAF analyses focus on the linguistic output of the participants. Because of this objective as well as the fact that the recordings were audio only, no information was added about background noises or non-verbal behavior, albeit laughing and chuckling were included to enhance readability of the transcript and to transport some of the participants' emphasis and of the interview mood in general.

At the time of the recordings, no automatic transcription tools were available that were able to reliably recognize spoken speech in general and L2 speech including the features of foreign accents in particular. Thus, multiple rounds of listening and manual transcribing were done to reliably capture all parts of the utterances that were necessary for the analyses. Those required parts included repetitions, self-repairs, false starts, and hesitation markers (*uh, uhm*).

Several notation systems offer means to represent features of spoken language such as rate of speech, volume, overlapping speech, or hesitations (e.g. Du Bois, 1991; Gumperz & Berenz, 1993). The Jefferson notation system (Jefferson, 2004) is a set of symbols used to represent speech. A modified Jefferson notation system was used for the transcription of the data as shown in the following section.

3.2.2.1.1 Transcription Conventions

As is common practice in deciding on what to include in transcriptions, the main guideline was to include as many features as necessary to approach the analysis validly and as few as possible so as to remain within given time restrictions. Estimates on the time required for transcriptions range from to 4 to 28 times the length of speech, depending on the transcription convention and level of detail (Nagy & Sharma 2013, p. 251). Some features, such as degrees of volume, were included to improve readability and to transfer some of the authenticity of the recordings into the transcripts. The

transcription conventions were adapted from Dressler and Kreuz (2000) and Jefferson (2004).

The following symbols were used for the interview transcriptions:

↓	A downward arrow indicates falling intonation.
↑	An upward arrow indicates rising intonation.
-	A dash indicates a cut-off.
" "	Double quotation marks indicate a shift in the speaker's voice when quoting.
° °	Degree signs indicate softer volume of the utterances spoken.
CAPITALS	Capitals indicate volume increase.
Underlining	Underlining indicates emphasis.
(.)	A micropause.
(0.6)	A number between parentheses indicates a timed pause.
[ ]	Square brackets indicate overlapping speech.
( )	Empty parentheses indicate an unclear utterance.
(word)	Filled parentheses indicate a likely, but uncertain word.
((laugh))	Double parentheses indicate aspects such as laughter.
italics	A word set in italics indicate a non-English word.
((TR:))	Translation of the non-English word into English
I:	Interviewer
T:	Teacher

CAF coding symbols in part 2 of interview:

//	AS-unit

::	Clause
{ }	Repair, repetition

3.2.2.1.2 Orthography and Raw Data Trimming

No punctuation marks were used to indicate sentence boundaries or other boundaries since segmentation in spoken language is done by intonation and pausing. Both may or may not coincide with common syntactical boundaries in written language. The utterances are transcribed in a run-on fashion, with pauses and intonation shifts indicating boundaries in the string of speech.

The spelling in the transcripts follows standard orthography, regardless of accents, unless there is a difference in meaning or grammar. That was the case in words such as *spent,* which could have been the simple past form of *spend*, but could have just as well been a case of final devoicing, in which the voiced alveolar plosive [d] was replaced by its unvoiced counterpart [t]. Devoicing final consonants, which would be voiced in standard forms of English, is a very frequent phenomenon observed in speakers for whom English is a foreign or second language and whose first language does not entail particular voiced consonants in final positions, German being one of those languages (see e.g. Brockhaus, 2012). Thus, the word *spent* in the participants' spoken English could have been a matter of mispronouncing *spend* rather than a deliberate choice of tense. In order to keep this ambiguity transparent, the transcripts represent the word as the actual spoken variant in its orthographic form. Capitalization was done for the pronoun *I* and proper nouns. Pause fillers were transcribed as orthographic approximates, such as *uhm* and *uh*.

Because some of the fluency measures (see section 3.2.2.2.4) were calculated in the computer application PRAAT (Boersma & Weenink, 2005), the recordings needed to be converted into an applicable format for the computer application. The raw interview recordings were trimmed in the computer application Audacity ("Audacity (R): Free Audio Editor and Recorder [Computer program]," 2015) so as to isolate the monologic second part of the interviews. The monologic interview parts present the core data as they recorded uninterrupted continuous speech. They were isolated from the complete interviews in order to run fluency scripts in PRAAT. Because there was only one speaker and only little ambient noise, the PRAAT fluency script was highly reliable to detect the interviewees' speech.

3.2.2.2 Coding

The transcripts were coded line-by-line. First, the speech was segmented into AS-units. Next, the coding for the corresponding measures of each CAF dimension was done either on the transcripts or the recordings, or on both. The manner of coding the data greatly determines the analysis results. Therefore, the following sections give a detailed account of the coding procedures.

3.2.2.2.1 AS-Units

In order to segment the transcribed speech, AS-units of the participants' speech in the interviews were determined according to the definition by Foster et al. (2000): "An AS-unit is a single speaker's utterance consisting of an *independent clause, or sub-clausal unit,* together with any *subordinate clause(s)* associated with either" (p. 365). A subordinate clause in this sense "will consist minimally of a finite or non-finite verb element plus at least one other clause element (Subject, Object, Complement or Adverbial)" (Foster et al., 2000, p. 366).

In line with Foster et al.'s definition of an AS-unit, the following example of Teacher 1 is counted as two clauses and one AS-unit:

(1) // so he challenged us :: to uhm try our best // (Appendix A)

In the sentence *I like reading*, the word *reading* is considered a noun phrase because according to Foster et al. (2000), at least one additional clause element is needed to receive clausal status.

Coordinated verb phrases usually constitute an AS-unit, unless the first phrase is followed by a rising or falling intonation and a minimum pause of 0.5 seconds. Foster et al. (2000, p. 367) argue that a pause of such a length is clearly noticeable and can be determined without any specific equipment. One of their examples is the following utterance:

> `The other woman is very happy now (0.5) and (3.0) <u>just walking away with a gr great smile</u>.` (Foster et al., 2000, p. 364)

In this example the second verb phrase indicates a new beginning for the speaker even though the subject is missing. Subject-dropping as in the second phrase is a common feature of non-native speech although probably not intended to be coordinate phrases (Foster et al., 2000, p. 364). Accordingly, similar cases in the transcripts of the present study were segmented as two AS-units as in the following example in Teacher 4's interview:

(2) //she just saw what I could do↑ and (0.92) //showed me a way to get better (Appendix D)

Seedhouse, Harris, Naeb, and Üstünel (2014) adapted the definition to reach high inter-rater reliability and to accommodate for features of speech. In their study on the relationship of quantitative measures and IELTS descriptor bands, Seedhouse et al. (2014) identify boundaries of AS and A units, which is the term they use for subordinate clauses, with at least a finite or non-finite verb element, as follows:

> As a rule, the existence of falling or rising intonation followed by (0.5) pause identifies the start of a new AS unit. When there are cases of doubt, count the utterance as a separate A unit to show complexity. (p. 26)

Thus, pause length was included in setting unit boundaries. In addition to pause length, Seedhouse et al. (2014) include falling intonation in their working definition of an AS unit: "A noun phrase without a verb is considered a separate AS unit if it is separated from the following phrase by falling intonation and a pause of (0.5)" (p. 26). This definition implies that falling intonation indicates a boundary in the planning process.

In the segmentation of the interviews of the study at hand, not only falling but also rising intonation was included as a possible boundary between units as well. High-rise terminal refers to rising intonation at the end of a statement and is a common feature in some varieties of English, such as New Zealand English, Australian, Northern Ireland English, Wales and Northeast England (Crystal, 2018b). In addition, high-rise terminal, also called *uptalk,* is heard in North American English and is continuously spreading particularly with teenage and young adult English speakers (Warren, 2016). As some of the interview participants engaged in high-rise terminal intonation in their speaking as well, rising intonation followed by a pause as well as falling intonation indicated a boundary between two AS-units. Rising intonation followed by a pause equal or longer than 0.5s was considered a boundary as well as falling intonation followed by a 0.5s or longer pause as in the following example by Teacher 3:

(3) I think :: I'm now (1.42) very strict↑ :: (0.57) // {to} to pupils↑ (Appendix C)

Subordinate clauses were normally coded as part of those AS-units, which included the corresponding independent clause. However, adjustments needed to be made to code subordinate clauses that were separated from their independent clause by rising or falling intonation plus a pause of 0.5 seconds or longer. This was sometimes the case with subordinators or

coordinators such as *because* and *but,* both of which can serve as part of an ellipted independent clause and as such are quite frequent in spoken language.

Because the segmentation into AS-units would influence those CAF ratios, which are based on AS-unit segmentation and in order to achieve high intra-rater reliability, segmentation was done at least three times. An additional rater, who was experienced in this type of coding, was asked to segment some of the transcript samples (9.09%) according to the given definition of an AS-unit as well. The researcher herself coded all of the interviews so that consistency in the coding was guaranteed.

Pauses of 0.5s or longer as well as noticeable rising or falling intonation were included in the transcripts and determined those AS-unit boundaries which would otherwise not be drawn, as shown in the following utterances by Teacher 1:

(4) // and he said "I (0.41) expect this and that and that from you"↑ (0.82) // but on the other hand he was still fair and nice and friendly and we still liked him↓ // (Appendix A)

3.2.2.2.2 Complexity

In order to tap into complexity, three measures were applied – two for syntactic complexity and one for lexical complexity. The measures for syntactic complexity were based on subordination and length of AS-unit.

Subordination was counted in the number of subordinate clauses. A subordinate clause "will consist minimally of a finite or non-finite Verb element plus at least one other clause element (Subject, Object, Complement or Adverbial)" (Foster et al., 2000, p. 366). The measure was the ratio of subordinate clauses to AS-units.

In those cases where the subordinate clause was not included in the AS-unit of which the independent clause was a part because of rising or falling intonation and a pause equal or longer than 0.5s, the clause was also not counted as a subordinate clause.

Clauses with *because* were evaluated according to the pausing and intonation pattern. When *because* followed the independent clause at the end of the according AS-unit with the same intonation without a 0.5s pause, the *because*-clause was considered a subordinate clause. It was not considered a subordinate clause when it introduced a new complex utterance not directly related to the preceding independent clause.[17] That way, the particular function of *because* in spoken language was taken into

17 For an extensive discussion of *because* in spoken language see Chafe (1988).

consideration. The following example of Teacher 8 illustrates how *because* can introduce a new unit:

(5) // and uhm so it was just a different way of (0.43) teaching↓ :: // because we went into class :: she said (0.27) :: "well everybody put on your jackets :: // we go outside"↓:: // (Appendix H)

An additional measure was computed for the number of clauses in total. As studies are not clear on whether they considered subordination being based on all clauses including coordinate clauses or merely on subordinate clauses, two measures were included here – one based on all clauses and a second one based on subordinate clauses only.

Lexical complexity was measured in lexical diversity, namely vocd. Since the value D is calculated by random sampling, as explained in section 2.1.4.2 on operationalizing complexity, and therefore results in different values each time, the calculation was done three times, using the online tool Text Inspector (2016). According to McCarthy & Jarvis (2010, p. 383), a three-fold calculation creates a higher level of consistency and results in an average D as the final output. The average of those three values obtained in the Text Inspector calculations was then used as the score for lexical complexity (Appendix L).

In line with Foster et al.'s (2000) approach, false starts, repetitions, self-corrections were excluded for measures that included calculating words per unit for complexity, as repair phenomena and hesitations would inflate the number of words and result in confounded indices.

3.2.2.2.3 Accuracy

Accuracy was measured in the percentage of error-free clauses. Clauses that had errors in syntax, morphology, or lexical choice were not considered error-free. In line with Yuan and Ellis (2003), lexical errors were defined as such if the clause showed errors in lexical choice or collocation, for example in "'I was waiting you`" (Yuan & Ellis, 2003, p. 13).

As the focus of this study is to account for overall accuracy, the measure was a general measure, namely the ratio of error-free clauses. Clauses containing errors which were self-corrected by the speaker immediately after the error occurred were considered correct (R. Ellis & Barkhuizen, 2005). If the corrected elements resulted in an erroneous clause, the clause was not considered correct.

In the utterance "so he decides decided decided to go fishing" (Foster et al., 2000, p. 368), the clause would be considered correct. Accordingly, the following sample clause by Teacher 3 was counted correct because there

was an immediate correct self-repair changing the incorrect present tense in *I think* to the past tense obligatory in the given context *I thought:*

(6) // and then uh (1.37) {I think} (0.35) I thought :: // (Appendix C)

The following examples, each of a different teacher, illustrate clauses that were not coded error-free:

(7) and after I start in the ninth (Teacher 3, Appendix C)
(8) who came new to that school (Teacher 2, Appendix B)
(9) he he didn't taught taught us any theory (Teacher 4, Appendix D)
(10) what what me influenced (Teacher 6, Appendix F)
(11) and trying to watch out for seals [context meaning was *watch seals*] (Teacher 7, Appendix G)
(12) and I think movement would be great [context meaning was *exercise*] (Teacher 8, Appendix H)
(13) to to show us what the culture in English like (Teacher 9, Appendix I)
(14) because we do that at sport as well (Teacher 10, Appendix J)
(15) I want to become teacher (Teacher 11, Appendix K)

3.2.2.2.4 Fluency

The following fluency measures were applied based on studies measuring fluency (e.g. Bosker, Pinget, Quené, Sanders, & de Jong, 2012; de Jong & Bosker, 2013; de Jong, Groenhout, Schoonen, & Hulstijn, 2015; Kormos & Dénes, 2004; Tavakoli, 2016; Tavakoli, Campbell, & McCormack, 2016):

Table 2 *Fluency measures*

Subdimension of fluency Measure	Method
Speed fluency	
Syllables	PRAAT Script Syllable Nuclei v2 (N. H. de Jong & Wempe, 2009). Pause threshold of 0.25 sec. Silence thresholds from -25dB to 40dB, depending on the interview audio quality.
Speechrate. Mean length of syllables	Number of syllables divided by total time

Subdimension of fluency Measure	Method
Mean length of runs	Number of syllables divided by number of runs. Run indicates utterances between pauses of 0.25 second and above.
Articulation rate	Number of syllables divided by phonation time
Breakdown fluency	
Silent pauses Total number of silent pauses	Silent pause threshold 0.25s. PRAAT application with different silence thresholds depending on quality of audio files.
Silent Pause Duration	Total time of all silent pauses. All timed pauses were included in transcripts, shortened to two decimals and added up.
Mean Pause Duration	Total silent pause duration divided by the number of silent pauses
Silent pauses per minute	Number of silent pauses divided by total recording time times 60
Phonation Time	Total recording time minus silent pause duration
Phonation time ratio	Phonation time divided by total time
Filled pauses. Number of filled pauses per phonation time	*uhms*, *uhs*, and fillers such as *yeah* and *like*. Count on transcripts.
Filled pauses per minute total duration	Number of filled pauses divided by total time times 60
Repair fluency	
Repairs. Number of self-repairs	Instances of self-repairs. Count on transcripts.
Repairs per minute	Number of repairs divided by recorded total time times 60.
Repairs per speaking time	Number of repairs divided by recorded total time.
Repetitions. Number of repetitions	Repeated words or phrases. Count on transcripts.

Subdimension of fluency Measure	Method
Repetitions per speaking time	Number of repetitions divided by phonation time
Repetitions per minute	Number of repetitions divided by total time times 60

Silent pauses were measured in the software application PRAAT (Boersma & Weenink, 2005). The pause threshold was 0.25s. The cut-off point of pauses has been discussed to correlate with proficiency levels when it is set between 0.25s and 0.3s (N. H. de Jong & Bosker, 2013; Segalowitz, 2016). In addition, Towell et al. (1996) suggest a pause threshold of 0.25s because a lower threshold may include plosives or other phenomena that are not hesitations.

The latest PRAAT script developed by N. H. de Jong and Wempe (2009), PRAAT Script Syllable Nuclei v2 2017, was applied to measure the syllables in each interview recording. The script also measures pauses, articulation rate, and phonation time. The script was developed to determine features such as speech rate in large samples in a reliable and automatic way. In order to do so, however, the audio has to be of high audio quality without background noise.

To a great extent, the interview audios at hand were of good sound quality and little background noise. The decibel (dB) silent thresholds, which determine at what volume PRAAT will consider a sound as a sound, were fixed according to the individual audio quality. They turned out to be most reliable between –25dB and –40dB. However, all the script calculations were also hand-edited since the PRAAT script occasionally mislabeled sounds for pauses, when for example an [s]–sound at the end of the word was elongated but too soft to be picked up by PRAAT as sound. In addition, the pauses were annotated by the PRAAT silences text grid, in which the output can be manually checked while listening to the original recording (*Figure 4*).

As it showed while computing the pauses in the Syllable Nuclei Script v2 (N. H. de Jong & Wempe, 2009) and in the built-in *Annotate Silences* function in PRAAT (Boersma & Weenink, 2005), the pause calculations were not exactly the same in the two outcomes. Personal communication with Nivja de Jong, David Weenik, and Ton Wempe (August 17, 21, 22, and 27, 2018), revealed that this was due to the distinct ways both commands are programmed: The syllable script detects syllables of normal spoken sentences and takes into account the variations in the amplitude within one syllable. The amplitude threshold is not held on one level but depends

on the directly preceding amplitude contour, so that the syllables can be accurately detected.

The phonation times computed in the Syllable Nuclei Script v2 for each audio were close to the manual calculations. In order to keep internal consistency in the calculations, all measures were applied correspondingly for each audio by a combination of automatic measurements and manual calculations.

For the number of filled pauses, hesitation markers such as *uhm* and *uh* were counted in the interview transcripts. In cases such as "and uh (2.28) uhm (1.56)" (Appendix C), silent pauses are the timed silences in the parentheses, the other *uh* and *uhm* are counted as filled pauses.

As all the interviews except one took place in a quiet surrounding, the recording quality was high with only little background noise. When there was humming, the dB threshold was lifted so as to capture the silences correctly. The PRAAT results were double-checked to ensure that the silences were placed correctly (*Figure 4*).

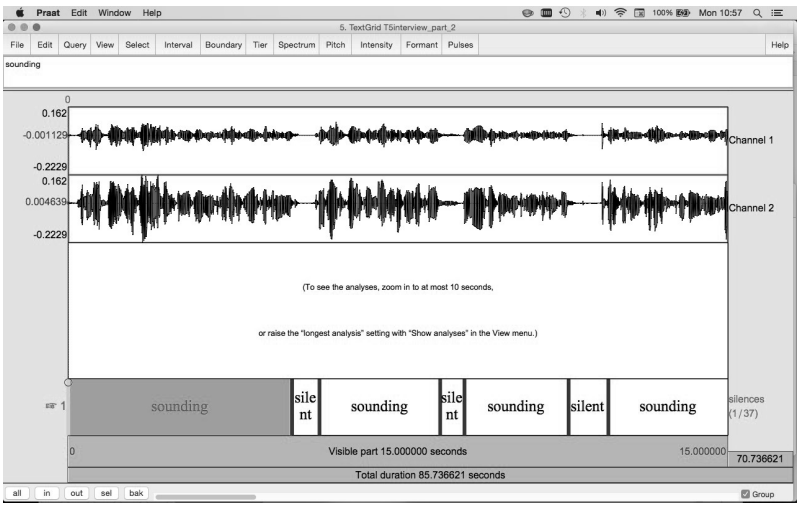

Figure 4 Screenshot PRAAT silent pauses

When the speaker repeated the same word or words, this was considered a repetition. Repetitions used for emphasis were not included in the count, as they cannot be considered dysfluencies as for example the following utterance shows: "it's a very very bad man`" (Foster et al., 2000, p. 368). Accordingly, *really* in the following utterance by Teacher 2 was considered emphasis and not coded as a repetition:

(16) she was really really cool↑ (Appendix B)

A false start indicates a part in an utterance that is either abandoned entirely or reformulated immediately after (Foster et al., 2000). The following utterance by Teacher 6 illustrates how repetitions and false starts were handled:

(17) // so {I had} {I had} or {I} I knew↑ :: (0.66) children need (1.00) the positive things↓ :: (0.50) // (Appendix F)

I had I had is a repetition as well as a false start that is rephrased as *I knew. I knew* is the repair, also called reformulation or restart, because it replaces an abandoned previous phrase by introducing a new formulation. Thus, repairs may or may not follow errors. In this example *I I* is another repetition.

When transcripts needed to be cleared of self-repairs, hesitation markers, and repetitions, as was done for calculating lexical diversity, the final part of the repetition or repair was left in. Thus, in the above example the pruned utterance was '*so I knew children need the positive things.*' When a measure required pruned speech instead of verbatim speech, this was indicated. A self-repair was also coded when an error was immediately corrected by the speaker, for example in the following speech sample by Teacher 3:

(18) and the first time (0.48) {I don't} (0.48) I didn't like her uah↓ (Appendix C).

The phrase *I don't* started out with the wrong tense in the given context and was self-repaired by introducing *I didn't.*

3.2.3 Results of Interview Language Performance

The following sections report the results of the calculations for each dimension – complexity, accuracy, and fluency – for each teacher participant. The large number of measures requires a focus in this section. Therefore, the following sections are restricted to depicting some of the teachers' values on complexity, accuracy, and fluency in each dimension on selected measures as an overview (complete scores, see Appendix L). The selection is based on those measures that have been reported repeatedly in CAF studies presented in chapter 2.1. The following result reports therefore need to be considered being exemplary showcases at this point. The aim is to provide a general overview of the teachers' results as background to the sections following thereafter. A more comprehensive analysis based on a

Principal Component Analysis for each CAF dimension is carried out as part of the final synthesis study 3 in chapter 3.4.1.

3.2.3.1 Complexity

Since complexity entails two sub-dimensions, the following sections show the results for both syntactic and lexical complexity.

3.2.3.1.1 Syntactic Complexity

As there was some variance in the total amount of time the participants spoke during the second part of the interview, the absolute values such as number of AS-units, number of clauses, number of subordinate clauses were considered as the numerators and denominators for the calculated ratios *clauses per AS-unit* and *subordinate clauses per AS-unit*. Table 3 is shown here to illustrate how the measures were handled without speaking time confounding the results.

Table 3 *Results interviews syntactic complexity*

	AS-units	Minutes total	Number of clauses	Clauses per AS unit	Subordinate clauses	Subordinate clauses per AS unit	Subordinate clauses per recorded minute
Teacher 1	29	1.91	54	1.86	12	0.41	6.27
Teacher 2	28	2.48	54	1.93	12	0.43	4.83
Teacher 3	26	2.43	40	1.54	2	0.08	0.82
Teacher 4	53	3.35	88	1.66	23	0.43	6.87
Teacher 5	20	1.43	34	1.70	10	0.50	7.00
Teacher 6	26	1.98	34	1.31	7	0.27	3.54
Teacher 7	38	2.46	58	1.53	9	0.24	3.67
Teacher 8	52	2.93	95	1.83	12	0.23	4.09
Teacher 9	24	2.09	25	1.04	6	0.25	2.87
Teacher 10	22	1.68	29	1.32	6	0.27	3.58

	AS-units	Minutes total	Number of clauses	Clauses per AS unit	Subordinate clauses	Subordinate clauses per AS unit	Subordinate clauses per recorded minute
Teacher 11	25	2.01	25	1.00	5	0.20	2.48

Table 3 shows the values for each teacher and the measures for syntactic complexity. Teacher 5, for example, spoke the shortest and had the fewest AS-units, but had the highest number of subordinate clauses per AS-unit. Thus, Teacher 5 can be considered to perform comparably high on syntactic complexity.

Figure 5 exemplarily shows the ratios of subordinate clauses per AS-unit for every participant. The ratio of subordinate clauses per AS-unit is a frequently applied measure in studies on syntactic complexity (see section 2.1.4.2).

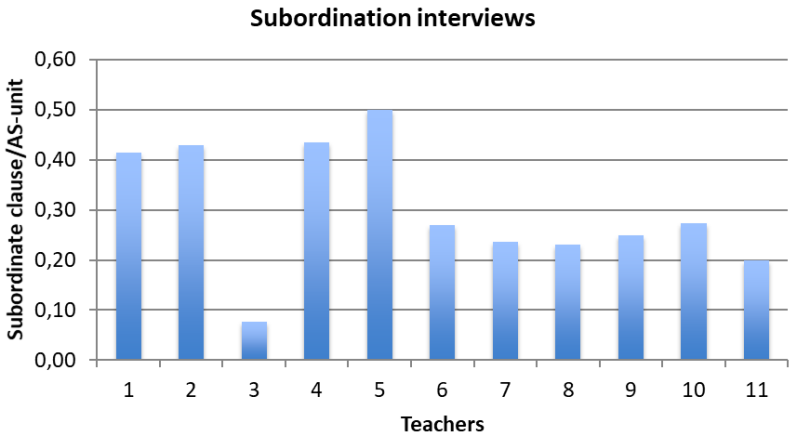

Figure 5 Syntactic complexity in interviews

The complexity range was from the minimum of 0.08 subordinate clauses per AS-unit of Teacher 3, to the maximum of 0.5 subordinate clauses per AS-unit by Teacher 5 (M = 0.301, SD = 0.127).

When the teachers' backgrounds were included for comparison, Teacher 3, who scored the lowest, did not hold a degree in English and had not lived in an English-speaking country. On the other end, Teacher 5, who

showed the most subordination in terms of subordinate clauses per AS-unit, had 5-7 years of teaching experience, held a degree in English and had lived in an English-speaking country.

In the subset of the four teachers whose students took part in Study 2, Teachers 1, 9, 10, and 11, Teacher 1 showed the most subordination followed by Teachers 10, 9, and 11.

3.2.3.1.2 Lexical Complexity

Because vocd is by definition based on text samples and therefore calculates slightly different results each time, the mean of three vocd calculations was computed and used as the lexical diversity measure of each teacher participant in the current study. The range of the participants' lexical diversity was from the minimum of 35.60, Teacher 3, to the maximum of 66.96, Teacher 7 (M = 53.18, SD = 10.83) (*Figure 6*).

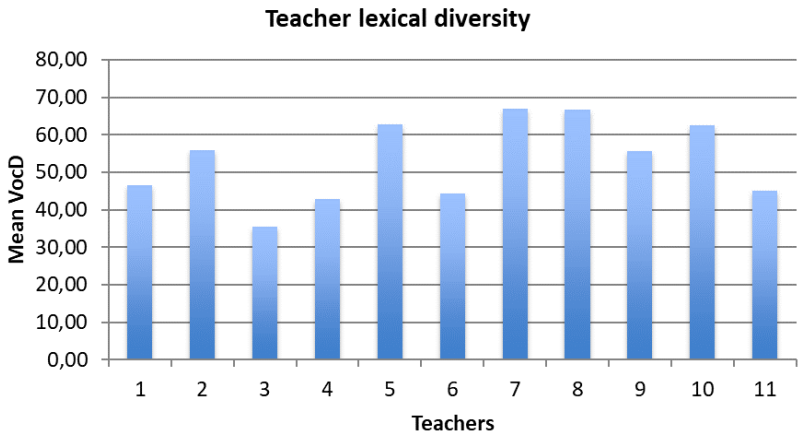

Figure 6 Lexical diversity in interviews

The results were varied with respect to the teachers' background: For example, Teacher 11, who did not have a degree in English and had not lived abroad in an English-speaking country, scored higher than Teachers 4 and 6, who both had a degree in English and had lived in an English-speaking country. The subset teachers showed some variance: Teachers 9 and 10 scored above the mean, while Teachers 1 and 11 scored below the mean.

3.2.3.2 Accuracy

The mean percentage of error-free clauses was *M* = 87.09 percent (*SD* = 10.82, *min* = 67.5, *max* = 100). *Figure 7* shows the percent of error-free clauses to total number of clauses.

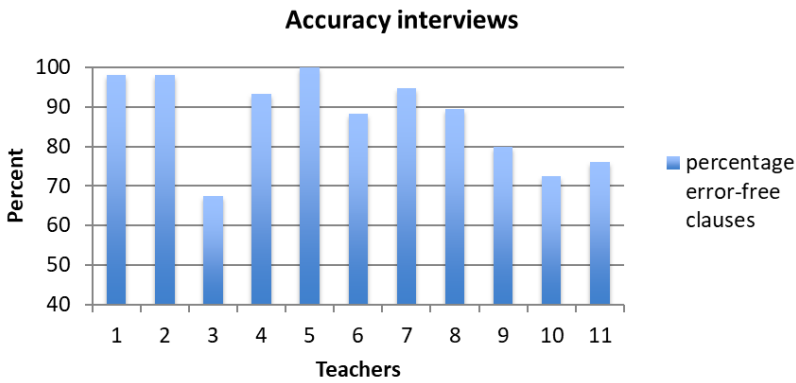

Figure 7 Accuracy in interviews

The ratio of error-free clauses to AS-units was from 0.76 to 1.89 (*M* = 1.34, *SD* = 0.39) (Appendix L). Of the teachers in the subset of four, Teacher 1 scored above the mean and higher on the percentage of error-free clauses than Teachers 9, 10, and 11, who scored below the mean, as illustrated in *Figure 7*.

3.2.3.3 Fluency

Table 4 shows the values for each of the fluency dimensions and the means of the total number of participants.

Table 4 *Fluency means in interviews*

	Minimum	Maximum	Mean	Std. Deviation
Mean pause duration (sec)	0.53	.91	.65	0.12
Number silent pauses	33	78	53.91	11.40

	Minimum	Maximum	Mean	Std. Deviation
Filled pauses per minute total dur.	2.10	16.50	8.46	3.83
Repairs per minute total dur.	0.00	2.53	1.56	0.73
Repetitions per minute total dur.	0.60	7.47	3.09	2.17
Mean length of runs	5.69	12.05	8.35	2.25
Speechrate syl./ total dur.	2.43	4.21	3.30	0.53
Articulation rate syl./ phon. time	3.50	5.33	4.48	0.56
Silent pause duration (sec)	17.59	51.23	35.13	9.70

Teachers $N = 11$

For two main reasons, fluency is broken down into three subdimensions: fluency is the dimension with the most measures, and a number of measures are phrased negatively. For example, a high value on the mean pause duration indicates less fluent language production whereas a high value on articulation rate would be indicative of more fluent speech.

Fluency was broken into the dimensions of speed fluency, breakdown fluency, and repair fluency. Speed fluency entails the speech rate and the articulation rate. Breakdown fluency includes the silent pause duration, mean pause duration, filled pauses per minute, and the mean length of runs. Repair fluency is measured in the ratio of repairs per minute and the ratio of repetitions per minute.

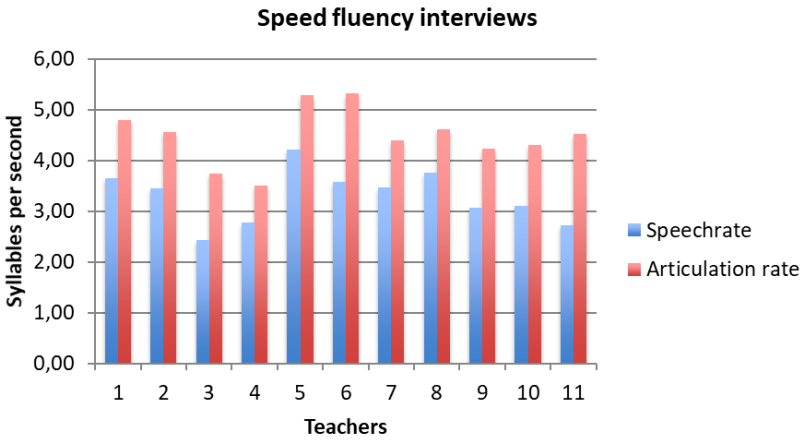

Figure 8 Speed fluency in interviews

In all cases, the articulation rate, calculated in syllables divided by phonation time, exceeded the speechrate, calculated in syllables divided by the total duration (*Figure 8*). The speechrate is always lower than the articulation rate because pauses impact the speechrate by using the total recording time as the denominator (M = 3.30, SD = 0.52). Pauses are excluded in the articulation rate. Articulation rate is measured in the number of syllables to phonation time, which is the actual time the speakers filled with speech (M = 4.48, SD = 0.56). To give an example: Teacher 5 scored high on both speechrate and articulation rate. This teacher spoke comparably fast during the interview, shown in speechrate. The speaker also produced a relatively little amount of pausing and therefore had a high articulation rate as well. Teacher 11, on the other hand, showed a greater gap between speechrate and articulation rate, which indicates that this teacher filled more time with pauses than Teacher 5.

Mean pause duration as one measure of breakdown fluency that is frequently used in studies on fluency performance (see section 2.1.4.4) is exemplarily illustrated in *Figure 9*. The mean pause duration shows how long the speaker's pauses were on average. The ratio is calculated by dividing the total silent pause duration by the number of silent pauses.

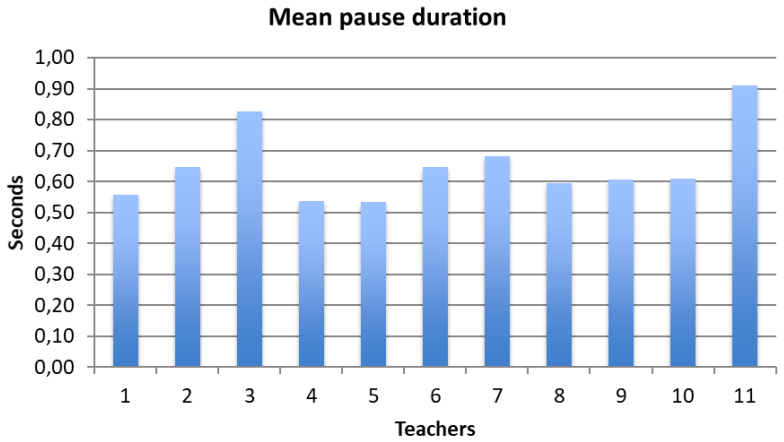

Figure 9 Breakdown fluency in interviews, mean pause duration

There is variance (M = 0.65, SD = 0.11) in the mean pause duration of the speakers ranging from the minimum of 0.53s in Teacher 5 to the maximum of 0.91s in Teacher 11. The difference between the two is 38 milliseconds, which can be a noticeable difference considering that the threshold for silent pauses was 0.25s. Of the four subset teachers, Teachers 9 and 10 had the same mean pause duration and both ranged between Teacher 1, who had the shortest, and Teacher 11, who had the longest mean pause duration.

Repair fluency per teacher, as measured in repairs per minute, ranged from minimum 0.00 to the maximum of 2.53 (M = 1.56, SD = 0.73). Repetitions per minute ranged from minimum 0.60 to maximum 7.47 (M = 3.09, SD = 2.17).

3.2.4 Discussion of Study 1: Teachers' Language Performance

Study 1 was guided by the research question of how the L2 English teachers performed in terms of complexity, fluency, and accuracy (RQ1). In the following, the findings of the interview study are presented and discussed. As the interview results are analyzed in more detail again in the synthesis study 3, the following section serves as a discussion on the most prevalent findings of the teachers' performance, their background, and their language use self-ratings in the questionnaire.

A number of measures for each dimension of language performance, complexity, accuracy, and fluency could capture features in the foreign

language production of the participants. The results showed that the teachers' performance differed from one another in all three dimensions – performance, complexity, accuracy, and fluency. The teachers on the lower end and higher end differed in all aspects. The teachers' syntactic complexity ranged from almost no subordination of 0.08 subordinate clauses per AS-unit of Teacher 3 to producing a subordinate clause in every other AS-unit of Teacher 5. The performance in lexical diversity showed a great range as well – a vocd score of about 36 to approximately 67.

The results on the accuracy percentage ranged from about 68 to 100 percent of error-free clauses. As was discussed in section 3.2.2.2.3 on accuracy, error-freeness is measured based on morpho-syntactic and lexical errors. Thus, an error-free clause was not equivalent to notions such as *elaborate* or even *perfect* and may not have been entirely adequate in all its dimensions, which explains the relatively high scores of all the participants on error-freeness. The results on fluency showed some variance in the lower performance and the higher performance as well.

In sum, the findings indicate that there was variance in the speakers' scores. It follows that the group of teacher participants in fact differed in terms of their language performance. The participants' self-selectivity as voluntary study subjects was expected to reflect their language performance, which would have resulted in similar scores among all participants. It was possible to include teachers in the study who scored comparably low on the language performance measured in this study as well as some teachers whose scores were comparably high.

The results also suggest that the participants may have scored similarly on each CAF dimension respectively: The participants performing comparably high on complexity also scored comparably high on accuracy and fluency and vice versa, at least in the selected measures shown in this section. The participants performing relatively low on one dimension also tended to score relatively low on the other dimensions. The relationships between the dimensions are the focus of Part I in Study 3 (section 3.4.1).

In terms of the teachers' background factors of teaching experience, degree status, and stay abroad, the results suggest two findings: At the top end of the CAF dimensions showcased in this chapter were teachers who held an English degree and those who had lived in an English-speaking country. On the low end was always a teacher without a degree in English and no stay abroad in an English-speaking country. Among the four teachers whose students took part in Study 2, Teacher 1 tended to score highest on the CAF dimensions, Teacher 11 lowest, and Teacher 9 and 10 alternatingly. The relations between the dimensions of complexity, accuracy, and fluency will be further analyzed and discussed in section 3.4.1 as part of the synthesis study 3.

Lexical diversity showed a slightly less systematic pattern among the teachers. Here too, the teacher scoring the highest (Teacher 7) held a degree in English and had stayed abroad, and the teacher scoring the lowest (Teacher 3) did not hold a degree in English and had not been abroad. However, Teacher 11, who was the other teacher without a degree, varied in her performance. Thus, further studies including teachers without a degree would be illuminating, yet difficult to engage as study participants, as was explained in section 2.2.5.

Regarding the subset of the four teachers, however, their performances varied more: Whereas in most other CAF dimensions shown in this chapter, Teachers 1 and 11 formed the top and low end, it was Teacher 10 who scored the highest on lexical diversity, but again Teacher 11 who scored the lowest.

However, because of the abundance of measures, it could not be determined from these results at this stage, how exactly the dimensions related to one another and which measures were most responsible for the relationships within each dimension. The relationships of the teachers' performance on each dimension is taken up again in Part I of the synthesis Study 3 (section 3.4.1), in which the measures form a composite score of each dimension based on a Principal Component Analysis.

Even though the focus of the present study was not the in-class language behavior of the participants, some information on how the teachers self-rated their classroom English use was considered to add to the discussion of the study findings, should the teachers report a diverging amount of English use in their classes. The four teachers, whose classes had been tested, were given a questionnaire to self-report their language use in the classroom. The questionnaire study is presented in the following sections.

3.2.5 Substudy: Teacher Questionnaire

While there has been research on language perception and beliefs of instructors as well as students in various instructional settings, little can be found on how observations on perceptions and beliefs relate to the language use in the classroom. A study done by Thompson (2009) with sixteen Spanish instruction classes in the U.S. attempted to bridged the two areas: questionnaires asked the teachers and students about the perceived teacher's target and non-target language choice. The results were then compared to classroom video-recordings. Thompson found significant correlations between the teachers' and students' perceptions of the teachers' language choice and the observed language use in the classroom. Concluding, those teachers and their students were able to reliably judge the teachers' actual classroom language choice.

In terms of self-efficacy, teachers' perceived efficacy was found to positively correlate with self-reported English proficiency (J. C. Richards, 2017). The same was reported in Eslami and Fatahi (2008), whose results showed a positive correlation between the teachers' perceived efficacy and their self-reported English proficiency.

A teacher questionnaire was developed for the present study following the research question (RQ2): How do the teachers rate their L2 English language proficiency and the modification of their language use in the classroom? No prediction was made for the self-rated language proficiency, but it was hypothesized that the teachers would report modifying their English language use in the classroom (H2).

The questionnaire included an item for language choice in the classroom as well as several phenomena of language modification (see section 3.2.5.1). The teachers' answers can be regarded as indicators of their actual language use in the classroom. However, questionnaires are limited in their quality criteria, as they cannot investigate an issue in depth due to the necessary simplicity of the questions (Dörnyei & Taguchi, 2010, p. 7). In addition, the reliability of the answers in questionnaires is limited because they are based on self-ratings. Reliability limitations in questionnaire data elicitation and self-ratings relate to self-deception and social desirability, for example, both of which may influence the participants' self-ratings (Dörnyei & Taguchi, 2010). By nature, the validity of questionnaires is limited as well, because the researcher cannot edit a participant's answers for their validity.

The merits of a questionnaire are its versatility in eliciting additional information. The questionnaires in the current study targeted information on the participants' perceived classroom language. The purpose of the questionnaire was to include quantifiable classroom language information to control for in later regression analyses, should the findings show significant relationships between teachers' CAF results and the students' test outcomes. The questionnaire answers were then analyzed for their moderating effects so as to include in-class target language use.

3.2.5.1 Data Elicitation Questionnaire

In order to gain additional information on the teachers' classroom language behavior, a teacher's questionnaire was developed to tap into the teachers' language behavior in the classroom (Appendix M). Even though the questionnaire could not quantitatively measure specifically how the teachers used English in the classroom, it could ask about the teachers' usage of the target language. The subset of those four teachers whose students were tested for receptive grammar and vocabulary in Study 2 were asked questions about their perceived usage of English. Teachers 1, 9, 10, and 11 filled out the questionnaire at their chosen time.

The items are explained in the following. Very little English in the classroom would have minimized the extent to which the children were exposed to the target language. Little exposure to English would have offered fewer opportunities for the students to actually hear a variety of words and structures. One question of the questionnaire was therefore an estimate of the amount of English used by the teachers in percent:[18]

(1) In class I speak ___% English with the children on average.

The possible answers were grouped into five ranges from 0–20% to 81–100%. The teachers were also asked to estimate their English language use in the classroom on a five-point Likert scale measuring language adaptation and rate their language proficiency.

Three sets of statements were developed to include estimates as to how much the teachers felt they modified their English in the classroom, how confident they felt about speaking English, and how they judged their own proficiency regarding teaching the language. On a five-point Likert scale, the participants were asked to answer from 1 (= strongly disagree) to 5 (= strongly agree).

Statement set (2) geared at how much the teachers modified their English as follows:

(2) When I speak English to my students, I…
(a) … slow down.
(b) … pause more often.
(c) … repeat words or phrases frequently.
(d) … simplify sentence structures by simple main clauses.
(e) … use simplified vocabulary.

The statements directly touch on two of the CAF dimensions, namely fluency (a, b), and complexity – syntactic (d) as well as lexical (e). Those five statements (2a-e) were integrated into a head category *adaptive language in class*.

No item was included prompting ratings about accuracy as answers were not expected to be reliable for mainly two reasons: The participants' reliable judgment of their own accuracy could not be guaranteed, and a halo effect was expected in the answers. Because high accuracy in the L2 is more respected than low accuracy, social desirability bias in the answers was expected to be strong.

Statement set (3) addressed how the speakers judged their own proficiency in English:

18 For the translated German questions see Appendix M.

(3)
(a) I'm confident with using English.
(b) I can always express myself in English.
(c) I have enough English skills to teach it.

The self-ratings to statements 3(a) and 3(b) were integrated in the head category *language proficiency speaking*. Statement 3(c) was headed *language proficiency teaching*.

3.2.5.2 Results of Questionnaires

All teachers reported speaking English in class most of the time. Three participants reported 81–100% English. One participant, Teacher 10, checked 61–80%.

Figure 10 shows each participant's rating means of each of the categories of *adaptive language in class, language proficiency teaching,* and *language proficiency speaking* (Appendix N).

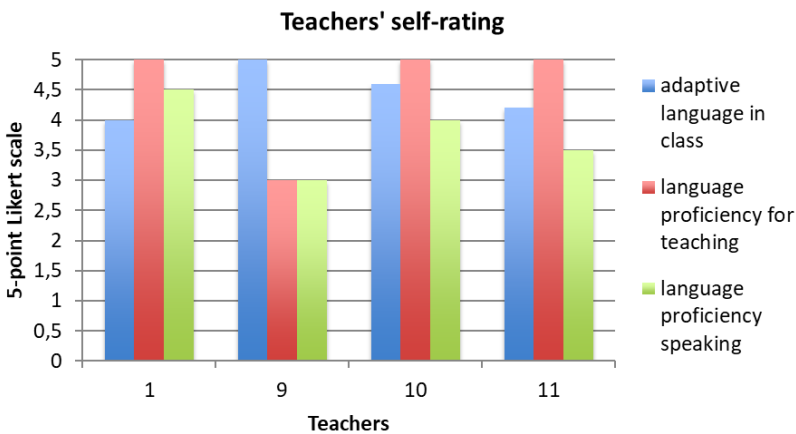

Figure 10 Teachers' self-ratings on adaptive language and proficiency

All four teachers reported adapting their language in the classroom (M = 4.45; min = 4; max = 5; SD = .38). On the self-rating of language proficiency for speaking, Teachers 1, 10, and 11 also rated their own language skills sufficient enough for teaching at 5, while Teacher 9 rated a neutral 3 for language proficiency for teaching (SD = .86). Teacher 9 also rated a 3 for general language speaking proficiency, while Teachers 1, 10, and 11 rated their general speaking proficiency at 4.5, 4, and 3.5 respectively (SD = .56).

3.2.5.3 Discussion Substudy: Teacher Questionnaire

(RQ2) How do the teachers rate their L2 English language proficiency and the modification of their language use in the classroom? (H2) The teachers report modifying their English language use in the classroom.

All four teachers reported speaking English in-class most of the time. The answers on language choice did not show any relevant variance, as three teachers, 1, 9, and 11, checked the highest option: 81–100% of English use in the classroom. One participant, Teacher 10, chose the next option down, 61–80% of English use in the classroom. Language choice was therefore not a reliably distinguishing factor between the participants. Yet based on these self-reports, the amount of English used in the classroom can be regarded as high in all classes, considering the elementary level of the students.

The answers on adaptive language and proficiency were slightly more varied than on the amount of English used in the classroom. First, all participants reported modifying their language in the classroom to some degree, as was hypothesized. Hypothesis H2 was therefore confirmed. Second, the ratings on how much the teachers modify their language in class showed less variance (SD = .38) than the other two components – the self-ratings on whether the teachers considered their language proficiency sufficient for teaching (SD = .86) and on how they judged their English speaking proficiency in general (SD = .56). Teachers 1, 10, and 11 rated their language proficiency high enough for teaching, but lower in general. This suggests that they did not consider the same amount of confidence in speaking the language a prerequisite for teaching the language. It is noteworthy that Teacher 1 was the one who reported performing the least language modification in the classroom. Teacher 9, who rated the lowest on both proficiency scales, rated the highest on adaptive language. This teacher's answers indicated that perceived lower language proficiency was linked to more adaptive language in the classroom. The CAF scores in Study 1 showed that Teacher 1 was the teacher who scored comparably high, while Teacher 9 ranged alternatingly with Teacher 10 most often in the middle. Therefore, no direct connection can be made between scoring relatively high on the CAF measures and the perception of in-class language behavior.

This result may indicate that higher linguistic performance does not necessarily accompany more modification in the classroom input. Assumptions offered in section 2.1.5 argued that teachers would need some base language proficiency in order to be able to modify their language. While this may still be the case, the current results show that a higher linguistic performance may not relate to in-class language modification. As this study was considered a substudy, more studies would be needed to conduct research examining the specific relationships between language performance and in-class modification.

In sum, teachers 10 and 11 showed a similar pattern in the self-ratings on the three components of adaptive language in class, language proficiency for teaching, and language proficiency in speaking. Teachers 1 and 9 showed nearly opposite patterns in their self-ratings: Teacher 1 felt very confident in speaking the language and also reported adapting the in-class language, but less than Teacher 9, who reported to strongly adapt, but felt medium confident with speaking the language. Both had a degree in English and had spent time in an English-speaking country. A difference between Teacher 1 and 9 was experience, with Teacher 9 having taught for more than ten years, whereas Teacher 1 had been teaching for two to four years. On the other hand, Teacher 11, who did not hold a degree in English and had not stayed abroad, rated her language proficiency for teaching as well as her general language speaking proficiency higher than Teacher 9, who had a degree and had stayed abroad.

The ratings on the teacher questionnaire were able to showcase some of the teachers' self-judgments of their language confidence and classroom use of the target language. Bearing in mind the limited sample size, the limitations in the reliability of self-ratings as suggested by Dörnyei and Taguchi (2010) as well as the focus and scope of the present thesis, the findings of the questionnaire study were considered additional information to the quantified findings in studies 1 and 2 as well as in the synthesis Study 3. However, they were only referred to when they qualitatively added to the respective findings and were then indicated as such.

3.3 Study 2: Students' Receptive Grammar and Vocabulary

In order to determine whether the teachers' language performances relate in any way to their students' acquisition of English, their students' English needed to be evaluated. The elementary schools did not have any standardized assessments and in fact, rarely assessed their students' English formally at all. Thus, there was no reliable evaluation of the students' development of English and no baseline data to start from. For this reason, two test instruments were chosen to evaluate the students' English, namely the BVPS3 (Dunn et al., 2009) and the ELIAS Grammar Test II (Kersten, Piske, et al., n.d.). Study 2 is motivated by the following research questions and hypotheses:

(RQ3) How do the students' receptive English grammar and vocabulary develop over their fourth year of elementary school? (H3) The mean scores in grammar and vocabulary increase between the two test times.

(RQ4) How do the student groups differ per teacher in their receptive English vocabulary or grammar attainment and development?

The purpose of the following study was to elicit an indication of students' language development of a subset of teachers over a period of time. In the state of Lower-Saxony, elementary school covers the initial four years of schooling. Therefore, the last year of elementary school was feasible for the times of testing, because the children would have been exposed to the target language for enough time to be able to show some initial language development. In addition, a comparably young age limited the extra-curricular influence of English media, which would have been a confounding factor to the teachers' language performance as an input factor.

In order to assess the students' receptive development of grammar or vocabulary, a longitudinal study was carried out in the classrooms. The learning environment of the participants was largely restricted to the classroom with their teachers as the main, if not only source of English, based on the students' self-reports. Before the testing, the students were asked whether they had contact with English speakers outside of the classroom or regularly visited English-speaking countries. Sources of English other than the teacher's input were virtually non-existent – neither in the children's personal environment nor in the larger context of German media and possible extra-curricular media influence. In particular, German TV-broadcasting of sources originally in English is dubbed and does not make an extra source of English. An English input source that is not controlled for may have been internet games.[19] However, the language input through internet games at this age may be considered limited to basic functional commands in the games. Overall, the conditions of the current study's participants' exposure to English can be considered being to a large extent controlled, with limited confounding factors to the teachers' language as an input factor. Hence, the children's learning situation was favorable for an analysis of the relationships between the teachers' language and the students' outcomes.

A limitation of test-based results in two areas of language acquisition is that no claims can be made about other language areas. In particular, contextual comprehension, such as in listening activities and reading comprehension, could illuminate a broader sense of general second language comprehension. In addition, specific limitations apply to the tests used in the study, which are described in section 3.3.1.2.

19 According to a survey on the number of children having smartphones, the vast majority of children between eight and nine did not have a smartphone: 18% of 8 to 9-year-olds had a smartphone in Germany in 2017 (Statista, 2019). The numbers increase in fifth grade, when students enter secondary school in many German states.

3.3.1 Data Elicitation

Twenty-five random regular elementary schools in the German state of Lower-Saxony were contacted and asked for participation. A fourfold consent match needed to be achieved in order to administer and use any testing at the schools: the respective teachers, the principals, the state board, and the parents needed to agree. This match was reached for four schools, four of the teachers interviewed in Study 1, and 132 of the students' parents. In total, the researcher administered 304 tests.

All children were asked by the researcher if they had English-speaking relatives, if they had lived in an English-speaking country or if they regularly spent their vacation in an English-speaking country. Their home languages were indicated on the answer sheets as well. The teachers supplied additional information on the home languages of the students.

The students were given a receptive vocabulary test (BPVS3), or a test of receptive grammar (ELIAS Grammar Test II), or both at two times. Time 1 was the end of grade three or the beginning of grade four. Time 2 was towards the end of grade four.

3.3.1.1 Participants

The participating classes and children were selected based on the consent matches given by the school principals, the teachers, the children's parents, and the school board. 132 regular public elementary school students took part at two test times between the mean age of 9.5 years (M = 114.4 months) at the first testing and 10.4 years (M = 124.7 months) at the last testing during the second half of the fourth grade. The number of girls to boys was 69 (52.3%) to 61 (46.2%) with two unreported genders (Appendix O, P).

The children were learners of English as a foreign language predominantly with German as their first or one of their first languages. None of the schools had extra-curricular English activities. Participants who had English as a first or one of the first languages were excluded from the analysis (n = 1)[20], as well as students who had extracurricular tutoring in English (n = 1). None of the remaining children reported having any English-speaking relatives, spending regular vacation time in an English-speaking country, or having lived in an English-speaking country.

At the time of the second testing in grade four, the children had been taught by the same English teacher throughout the school year at the mandatory two-lesson load per week for English instruction at regular

20 This student reported having an English-speaking parent, but had limited productive skills. The child reported that the home language was German and that the English-speaking parent did not live with the child.

elementary schools in the state of Lower-Saxony. The participating students were students of four of the interviewed teachers of four different schools: teachers number 1, 9, 10, and 11 (Table 5).

Table 5 *Number of participants per teacher*

Teacher	Total N	BPVS3 at t1	BPVS3 at t2	ELIAS II at t1	ELIAS II at t2
1	14	–	–	13	14
9	43	18	21	34	40
10	43	17	19	15	20
11	32	16	15	32	30
Total	132	51	55	94	104

Note. Total *N* = individuals. Some students took both tests at both times. The total number of tests administered is 304. Nine classes were tested.

At time 1 of the BPVS3 test, one group was about to finish their third year (Teacher 9). All the remaining groups were in the first half of their fourth year of elementary school at time 1 of both the grammar and vocabulary tests.

3.3.1.2 Test instruments

Since the participants had only had limited exposure to English and because they were still at an initial stage of language development, their productive skills in the target language were expected to be much lower than their receptive skills, making receptive tasks much more feasible and promising (Unsworth et al., 2015). Two domains of language were tested: receptive vocabulary and receptive grammar, each through standardized tests described in the following sections.

3.3.1.2.1 The British Picture Vocabulary Scale III (BPVS3)

The British Picture Vocabulary Scale (BPVS) is based on the North American Peabody Picture Vocabulary Test (PPVT), a standardized test that has been widely in use since its first publication (Dunn & Dunn, 1959). Both tests measure receptive vocabulary and have been revised several times. The BPVS is a later version than the PPVT and incorporates local words used in the U.K.

The test used in the present study was the third edition of the BPVS (Dunn et al., 2009), which includes English as an additional language (EAL) learners as possible test takers. The test contains four practice templates, followed by 14 sets of 12 test items, each successive set becoming more difficult than the preceding one. Each item shows four color pictures, of which only one corresponds to the prompt. The pictures represent words that would be found in everyday life contexts and increasingly less frequent words in the higher sets. An example is shown in *Figure 11*.

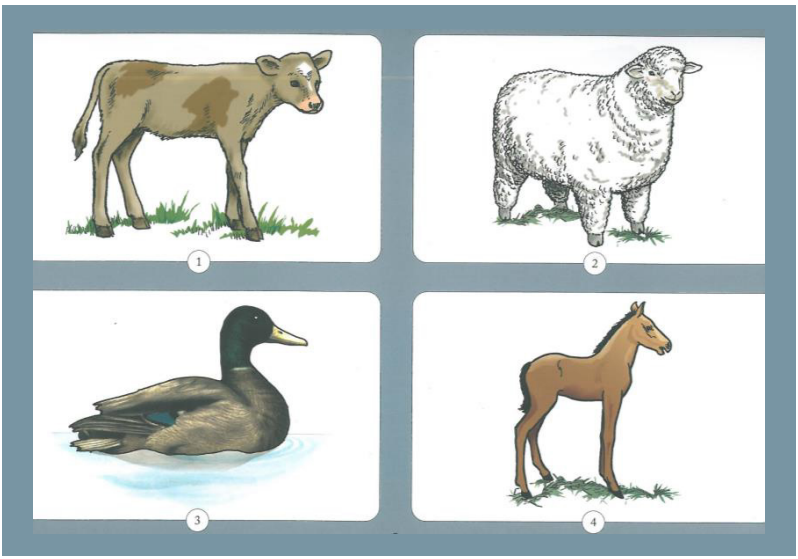

Figure 11 BPVS3 set 1 prompt 2: "duck"
(Dunn et al., 2009)

Based on abundant research on the validity and reliability of the PPVT, the BPVS is considered valid and reliable as well, but may have more predictive power if it is used next to other tests assessing different linguistic or cognitive areas (Harris, 2002).

Even though the test standardization included EAL learners, a limitation of the BPVS3 may be the validity of individual items and their applicability to a non-UK cultural setting. The BPVS3 includes some local variety words that may be troublesome for speakers in a non-UK environment. Two examples illustrate the difficulty of transferring culturally biased prompts to a different environment: *spanner* (set 4, item 47) and *waistcoat (*set 7, item 74). *Spanner* and the equivalent *wrench*, which is the standard word for example in North America and is also used in the PPVT, may possibly

be chosen by ruling out the distractors showing pliers, a file, and a chisel, but may be an invalid prompt if a test taker only recognized the word *wrench* but not *spanner*. An example of possible cultural as well as lexical appropriateness is the prompt *waistcoat* (set 7, item 74). The equivalent to *waistcoat* in other standard varieties of English is *vest*. The prompt *waistcoat* may be misleading particularly in a cultural context where this piece of garment is rarely seen. In addition, one of the distractor pictures in the *waistcoat* template shows a short jacket. Speakers not familiar with the term *waistcoat* may be tempted to choose the picture showing the short jacket, as *coat* and *jacket* refer to a similar referent of a garment for the upper body with a front opening ("coat," n.d.; "jacket," n.d.).

In order to avoid giving such a misguiding cue, the pronunciation [ˈwɛskɪt], which is also the preferred pronunciation according to the Merriam Webster Online Dictionary ("waistcoat," n.d.), was chosen over [ˈweɪstkoʊt] or [ˈweɪskoʊt] in the administration of the tests. *Vest*, the term used for the depicted garment in a North American environment, for example, would have given most students in the study a clue, as the German term *Weste* [ˈvɛstə] and the English word *vest* can almost be considered cognates, the only difference being the unstressed final Schwa in the German [ˈvɛstə].

In order to examine which term would be actively produced by English lecturers and future English teachers, the researcher carried out a small-scale survey with 15 German L1 speakers teaching or studying English. Thirteen of them were students in their third or fourth year of English studies, while two were lecturers of an English teaching program at a German university. The participants were shown the picture template and asked to say a word that would best describe the picture. The survey revealed that none of the participants produced the term *waistcoat.* Ten participants called the item in the picture *vest,* two produced *jacket*, one *shirt,* and four *don't know.* Nine participants had been abroad in English-speaking countries, namely in the UK, USA, and Australia, and one in Norway, where the study program was conducted in English.

The results of this survey exemplify the problem that particular words in the test may not be frequent in the children's environment, least so in an English-as-a-foreign-language setting.[21] The frequency of particular test words may additionally be affected by the L1 of the target language model speaker. For example, German has only one equivalent, *Schildkröte,* for two English words, *tortoise* and *turtle.* The word *tortoise* is a prompt in the

21 For a comparison of word frequencies in an English-speaking context, see Gnewuch (2014), who compared British children's spoken English listed in the CHILDES British English speech corpus to the frequencies of the words prompted in the BPVS.

BPVS3 already in the second set (item 16), but it may be the word less likely to be used by a German speaker of English than *turtle*. In fact, one of the teachers, whose students took part in the present study 2, commented after the test administration that she had never heard the word *tortoise*.

For the current study, however, the vocabulary test was intended to be altered as little as possible in order to maintain its reliability and the comparability of scores, both of which are strengths of the BPVS.

3.3.1.2.2 The ELIAS Grammar Test II

In order to assess the children's receptive grammar, the ELIAS Grammar Test II was chosen (Kersten, Piske, et al., n.d.). The ELIAS grammar test was first developed as part of the Early Language and Intercultural Acquisition Studies ELIAS project (Kersten et al., 2010).[22]

Like the BPVS, the ELIAS grammar test is a picture-pointing test of receptive English. The children are shown sets of three black–and–white pictures, one of which corresponds to a specific grammatical element. The first edition of the ELIAS grammar test prompts nine grammatical phenomena in 54 test items split between two parts A and B. The templates show three pictures – one distractor and two pictures that only differ in the grammatical item such as [+/- plural]. Of those three pictures the subjects are asked to choose the one correct one, as in the example in *Figure 12*.

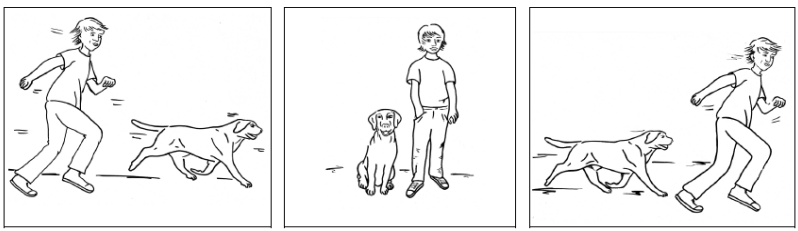

Figure 12 ELIAS Grammar Test II: "the dog is chased by the boy"
A set 1, prompt 4 (Kersten, Piske, et al., n.d.)

Nine grammatical phenomena are tested in the first edition of the test.[23] The ELIAS test has been revised once, adding relative clause constructions and passive mode prompts to the original edition and extending the number of

22 ELIAS was an EU Comenius project run between 2008 and 2010 and studied preschool children acquiring English as a foreign language in Sweden, Germany, Belgium, and the UK (Kersten, Frey, & Hähnert, 2008).

23 For a more detailed discussion of the grammatical phenomena in the test see Buyl and Housen (2015).

test items to 72 and twelve grammatical phenomena. The unpublished ELIAS Grammar Test II (Kersten, Piske, et al., n.d.) was used in the present study.

All items are based on contrasts and distributed evenly over parts A and B of the test. Each phenomenon is presented in several examples. When there are contrasting grammatical phenomena, such as plural versus singular forms, the test includes both grammatical forms with the same lexical item in the prompts of the A and B part respectively (e.g. *cat*, *cats*). By using the same lexical items for different grammatical phenomena, misleading wrong answers that could occur because of the unknown lexical items are limited. Yet the test takes a certain amount of lexical knowledge for granted and can therefore not entirely rule out that instead of – or in addition to – the grammatical phenomenon, lexical knowledge is tested as well. This limitation has also been acknowledged in Steinlen et al. (2010, p. 77). Presenting two pictures and prompts that contrast in the grammatical form plus a distractor point the test participants to the grammatical form so that even if they do not know the lexical item, they can identify the grammatical clue. Table 6 shows the grammatical phenomena and example prompts of the ELIAS Grammar Test II.

Table 6 *ELIAS Grammar Test II phenomena*

Abbreviation	Phenomenon	Example prompt
AGRc	Subject-verb agreement: copula verbs singular/plural	*the deer is white/the deer are white*
AGRv	Subject-verb agreement: full verbs singular/plural	*the sheep eats/the sheep eat*
GEN	Possessive case: absent/ present	*the girl is kissing the boy/the girl is kissing the boy's dog*
NEG	Sentences: affirmative/ negative	*the boy is running/the boy is not running*
PASS	Passive word order	*the dog is chased by the boy/ the boy is chased by the dog*
PLU	Inflectional morpheme: +/- plural –s	*cat/cats*
POSS	Possessive pronoun singular: masculine/feminine	*his cat/her cat*
PROog	Personal pronoun singular (object) masculine/feminine	*the girl is kissing him/the girl is kissing her*

Abbreviation	Phenomenon	Example prompt
PROsg	Personal pronoun singular (subject) masculine/feminine	*he is singing/she is singing*
REL I	Relative clause type I	*the boy who pushed the girl wore a hat/the boy pushed the girl who wore a hat*
REL II	Relative clause type II	*the boy who the girl is pushing is eating/the girl who the boy is pushing is eating*
SVO	Word order	*the boy is touching the girl/the girl is touching the boy*

Note. Adapted from Steinlen et al. (2010, p. 77) and extended by the additional phenomena in the ELIAS Grammar Test II (PASS, REL I, REL II).

Like the BPVS tests, the ELIAS grammar test is standardized. A limitation with the ELIAS grammar test is its comparably small monolingual benchmark group: the monolingual English group for the original ELIAS test was twenty English-speaking children in the UK (Steinlen, 2013).

3.3.1.2.3 Test Administration

In total, 304 tests were administered. The grammar test as well as the vocabulary test was administered as a group test during regular English class time. The merits of group testing over individually administered testing were its feasibility, timeliness, and only little interference in the teachers' schedules and students' class routines. Individual testing would have required numerous test days for each test time, taking the children out of their familiar classroom context, and an extra room to take the individual students. Staying with their peers in their familiar classroom, instead of with an unfamiliar test administrator in a separate room, might lower possible students' anxiety.

In order to test the reliability of group tests compared to individually administered tests, a study by Lüke, Ritterfeld, and Tröster (2016) focused on this aspect of testing. They found that the children scored similarly on the German version of TROG when tested individually and when in a group test situation. In addition, some studies in second language acquisition research have administered the BVPS and ELIAS Grammar Test as group tests, for example Hopp et al. (2018) as well as studies in the SMILE project.[24]

24 SMILE Studies on Multilingualism in Language Education (2014–2019). Niedersächsisches Vorab, VW Stiftung. University of Hildesheim.

With respect to group and individual testing in the vocabulary test, the second edition of the BPVS included scores of students who sat the test individually as well as in a group. The mean individual scores were found to be 3.2 marks higher than the group scores, with a correlation score of 0.75 (Lloyd M. Dunn & Dunn, 2009, p. 28). A correction factor was then calculated into the BPVS3 to adjust to the imputed group raw scores (Dunn & Dunn, 2009). However, more comparative research on individual versus group testing of the BPVS3 as well as the ELIAS Grammar Test II – the two tests administered in the current study – is lacking.

During the test administration in the classroom, the administrator gave instructions in English. The instructions were additionally given in German to avoid testing the understanding in the instructions' part and to enable all students to complete the task. The students were informed that they would not be graded and were encouraged to rather consider the tasks part of a game. As incentives, all children were promised small thank-you gifts for taking part in the tests.

The participants were given answer sheets with the picture templates of the respective test. The test templates were also shown on slides while simultaneously playing the respective words, which had been recorded by the researcher. The administration procedure was the same for both the BPVS3 as well as the ELIAS Grammar Test II.

Each prompt was played twice. The test administrator ensured that the students had physically heard the word or phrase and repeated it when necessary. The participants then marked the corresponding picture on the provided answer sheets. In accordance with the BPVS test administration manual, no cues of any sort were given (Dunn & Dunn, 2009). Nouns, for example, were presented without articles that would potentially give grammatical hints towards the objects as opposed to the activities in the picture templates. For all prompts, the same imperative word was played to the children, for example: "*Check duck*". The students were also encouraged to make a best guess choice about unfamiliar words or phrases if they were not sure which picture to check. One and only one picture was supposed to be checked for each prompt.

Neither of the two tests, the vocabulary nor the grammar test, was designed as a timed test, which allowed the test administrator to grant the individual students the time they needed to find the respective answers. Only after every student had chosen an answer and checked a box, the next picture template and word would be presented. A great deal of attention was paid to the pace of every individual student in order to overcome the limitation of the administration of the tests as group tests and to avoid time pressure.

During the test administration, a friendly and encouraging, non-testing atmosphere was created to keep the students motivated and to

prevent frustration. At both test times, the test started with the first set, as is recommended in the BPVS3 test guidelines for English as a foreign language speakers (Dunn & Dunn, 2009).

The intervals between the two times of testing had to meet two criteria: first, a possible practice effect needed to be minimized. The American Peabody Picture Vocabulary Test (PPVT), on which the British version is based, has a high test-retest reliability: 300+ subjects were retested on the PPVT-4 after an average interval of four weeks (Dunn & Dunn, 2007b). The reported reliability score was .91 to .94 (L. M. Dunn & Dunn, 2007, p. 55). Significant practice effects were not expected after the intervals in the present study, which exceeded the four-week interval in the PPVT reliability test by far ($M = 7.6$, $min = 7.2$, $max = 8.5$ months).

Test-retest reliability scores have not been reported for the ELIAS Grammar TEST II yet. Therefore, the time intervals between all vocabulary and grammar tests were intended to be as similar as possible in order to achieve comparability in the amount of instruction between test times. That way, the amount of instruction time was similar between student groups, and a possible effect of largely diverging time intervals would be as limited as possible.

Secondly, the test dates had to take into account limited time for administering the tests, as the dates had to be scheduled between school vacations as well as class trips and other school activities. School vacation slots at public schools in Germany vary each year. For the year of the post-tests, the summer vacation started in June, which left a comparably shorter time frame before the summer break than in school years lasting until July or August.

3.3.2 Data Analysis Students' Tests

Several scores were calculated for each student: a score for test time 1 and 2 for each respective test, and a score for the difference between the scores at time 1 and time 2, for each respective test as well (for raw scores, see Appendix O). The raw scores of the BPVS tests were computed according to the test manual (Dunn & Dunn, 2009): The raw score for each participant was calculated by determining the ceiling set, which was the set in which the participant had given eight or more incorrect answers. As all the participants were foreign language speakers of English, the basal was always the first set. For each student, the raw scores were calculated by subtracting the total number of incorrect responses from the number of the last item in the ceiling set. On the ELIAS grammar test, the correct answers in both sets A and B were counted, and the percentage of correct answers to total items

of both sets A and B was calculated for each participant. One total score per student was calculated for each test time.

In order to calculate a score that captures the development between test time 1 and 2, the crude gain score was calculated.[25] A crude gain score is the difference between two scores calculated by subtracting the score at time 2 from the score at time 1 (t2-t1). The calculation of the test scores was identical at both test times for each test respectively, which is a requirement for obtaining a crude gain score (Barnes et al., 1983, p. 69).

Analyses were performed for the entire group of students as well as for the students grouped with the respective teachers who had taught them. First, the results of the whole group of students informed about the grammar and vocabulary attainment at time 1 as well as time 2 in order to be able to calculate the development between the two test times. The group test scores were needed to answer research question RQ3, asking how the students' receptive English grammar and vocabulary developed over their fourth year of elementary school from time 1 to time 2. Hypothesis H3 was tested, which predicted that the grammar and vocabulary scores increased between the two test times.

Second, analyses based on each teacher's group of students could reveal differences between the groups taught by different teachers, which was asked in research question RQ4 about how the student groups differed per teacher in their receptive English vocabulary or grammar attainment and development.

A number of statistical procedures was carried out: A paired samples t-test with a factor *time* was conducted to calculate whether the mean differences at the two test times were significant for receptive grammar and receptive vocabulary – pair one being the BVPS raw scores and pair two the ELIAS grammar test percentages, each pair with time 1 and time 2 as variables.

The presentation of the results of the statistical analyses for the whole group follow more detailed descriptive statistics on the individual scores in each teacher group to illustrate the actual individual test results. As nine different classes were tested, the individual class results are shown to provide more insight into each class. Paired sample t-tests were run for the grammar and vocabulary results in each of the nine classes to analyze the possible significance within each group between the test times. By documenting the individual outcomes in each class, it was possible to add the differences between each teachers' group of students by also illustrating the amount of variability among the students, which would otherwise be subsumed in the group means.

25 For more on gain scores, see Barnes, Gutfreund, and Satterly (1983).

Next, four one-way ANOVAs were computed. Two ANOVAs with the vocabulary scores at time 1 as well as time 2 each as the dependent variable and teacher as the independent variable calculated whether the attainment in vocabulary was significantly different between the student groups sorted by teacher. Another two corresponding one-way ANOVAs were run with the grammar scores at time 1 and time 2 respectively. A post-hoc Tukey comparison then revealed between which of the teacher groups there were significant differences in the test results of the students.

Two additional one-way ANOVAs were computed to analyze significant differences between the development of each group based on their teachers: One ANOVA with the calculated vocabulary differences between time 1 and time 2 (t2-t1), the other one with the calculated grammar differences between time 1 and 2 as the dependent variable.

An additional repeated-measures ANOVA including both test times as the with-in subject variables, factor *time,* and teacher as the between-subject factor was computed. A repeated-measures ANOVA was expected to confirm the results of the paired samples t-test by including four levels on the independent variable for each of the four teacher groups. A repeated-measures ANOVA reveals whether the within-individual differences form a systematic pattern (Urdan, 2017). Thus, a repeated-measures ANOVA calculates the same as a paired samples t-test.

3.3.3 Results of Students' Tests

First, the results for all participants were computed to be able to analyze how the entire group of participants scored. Next, the student participants were grouped with their teachers according to the grammar and vocabulary tests they took. The group means per test were calculated as well as the total means of all participants in the respective tests.

Table 7 shows that 49 students took the vocabulary tests at both times and 87 students completed the grammar tests at both times. The number of participants who had completed a test at one of the times only was included when calculations were based on the results at one time only. This is indicated accordingly.

Paired-sample t-tests were computed to compare students' performances at the two test times, for the vocabulary tests and the grammar tests respectively (Table 8). The paired sample t-tests can only be computed based on those students who completed the grammar and vocabulary test at both times respectively.

Table 7 *Descriptive statistics of paired samples at both test times*

	Mean	*SD*	*SEM*
Raw score BPVS t1	45.76	13.759	1.966
Raw score BPVS t2	47.45	15.658	2.237
Percent ELIAS grammar test t1	51.07	5.550	.595
Percent ELIAS grammar test t2	54.39	7.616	.816

BPVS $n = 49$, paired
ELIAS $n = 87$, paired

The paired samples t-test showed no significant difference from time one to time two for the BPVS ($t(48) = .634$, $p = .529$), whereas the increase in the paired students' grammar scores from time one to time two was significant ($t(86) = 3.820$, $p < .001$) (Table 8).

Table 8 *Vocabulary and grammar, paired samples t-test (2-tailed)*

	Paired Differences						t	*df*	Sig.
					95% Confidence Interval of the Difference				
	Mean	*SD*	*SEM*		Lower	Upper			
BPVS t1 - BPVS t2	-1.694	18.697	2.671	-7.064	3.677	-.634	.634	48	.529
Grammar test t1 - Grammar test t2	-3.321	8.108	.869	-5.049	-1.592	-3.820	3.820	86	.000

Next, the data of those students who were taught by the same teacher were grouped together to analyze the student means separately for each student group by teacher. Table 9 shows the number of students per teacher and their mean scores on the respective tests at time 1 and time 2. Here, all students who completed the test at either time were included.

Table 9 *Mean test scores per teacher group and total*

Teacher		Raw score BPVS t1	Raw score BPVS t2	Percent ELIAS grammar test t1	Percent ELIAS grammar test t2
1	M			49.89 (n = 13)	49.11 (n = 14)
	SD			5.180	6.596
9	*M*	49.00 (n = 18)	49.33 (n = 21)	52.45 (n = 34)	56.04 (n = 40)
	SD	12.462	18.271	6.202	10.467
10	*M*	44.18 (n = 17)	45.79 (n = 19)	52.78 (n = 15)	52.15 (n = 20)
	SD	12.566	13.794	4.896	4.863
11	*M*	45.38 (n = 16)	45.47 (n = 15)	49.91 (n = 32)	56.62 (n = 30)
	SD	16.645	15.137	5.075	4.280
Total	*M*	46.25 (n = 51)	47.05 (n = 55)	51.29 (n = 94)	54.53 (n = 104)
	SD	13.816	15.804	5.575	7.984

Note. Teacher 1's students did not take the BPVS. *n* corresponds to test participants. Some students completed the grammar as well as the vocabulary tests. Included are also students who only took the test at one time.

The results in Table 9 can be summarized as follows: The mean BVPS raw score of the entire group of test takers at time 1 and time 2 was approximately 46 and 47 respectively. The mean percentage of correct answers on the grammar tests was approximately 51% at time 1 and 54% at time 2.

The mean grammar percentages in Teacher 1's group went down slightly from time one, approximately 50%, to approximately 49% at time two. Teacher 9's group mean of the BPVS raw scores stayed about the same from time one to time two: 49% at time one and approximately 49% at time two. The mean percentage of the grammar tests in teacher 9's group went up from approximately 53% at time one to about 56% at time two. Teacher 10's students' mean vocabulary score went up from about 44 to approximately 46, but their mean grammar percentage went down slightly from approximately 53% to about 52%. Teacher 11's students' mean vocabulary score went up slightly from approximately 45 to about 46, and their mean grammar percentage increased from approximately 50% to about 57%. The

significance of each of the differences is shown for each group of students in the following sections 3.3.3.1ff, in which the individual groups are analyzed in more detail.

The subject differences are presented in more detail per teacher in the following sections. They illustrate the test results of each individual participant per teacher group at both test times. The individual vocabulary scores per teacher group are presented as well as the grammar scores.

3.3.3.1. Classes Teacher 1 Results Grammar

Teacher 1 had lived in an English-speaking country and held a degree in English. The school was in a suburban area and had about 480 students. In the teacher questionnaire, Teacher 1 had an adaptive score of 4, a score of 5 for self-rated speaking language proficiency, a score of 4.5 on language proficiency for teaching, and reported speaking 81–100% English in class (Appendix N).

The group in this class was available for the grammar test only. 13 children took part at time one and 14 at time two. The interval between the test dates was 247 days or 8.1 months, excluding the test date at time two. In this group about 10 children did not take part in either test because their parents did not give their consents.

During test time one, a child dropped out after the first half, which was most likely due to a diagnosed attention deficit syndrome. The results of this child were excluded from analysis. Of the 13 students who completed the tests at both test times, the percentages of 6 students increased, 1 stayed unchanged, and 6 decreased (*Figure 13*).

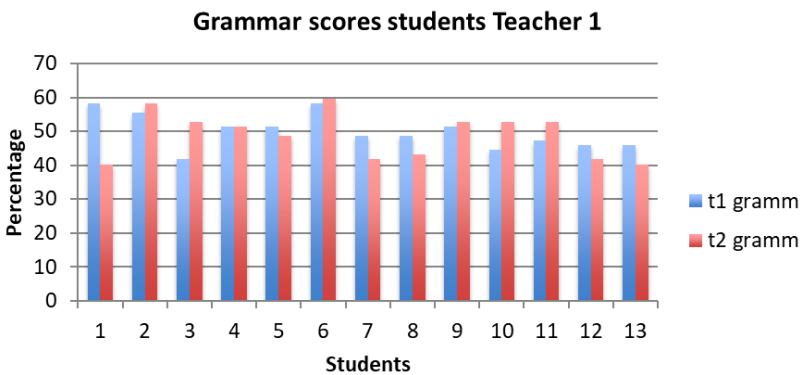

Figure 13 Grammar scores of Teacher 1's students at two times

The mean percentage for the grammar test at time one was 49.89 (*SD* = 4.98). The mean at time two was 49.11 (*SD* = 6.36). There was no significant difference between the means of the grammar test at time one and time two. A one-tailed paired samples t-test was not significant for the difference between the grammar scores at both times (t (12) = 0.316, p = .377). The mean grammar score for this group was the lowest of all the groups at time 2.

3.3.3.2 Classes Teacher 9 Results

Teacher 9 had lived in an English-speaking country and held a degree in English. The school was in a suburban area and had about 160 students. In the teacher questionnaire, Teacher 9 had an adaptive score of 5, score of 3 on both speaking language proficiency and language proficiency for teaching, and reported speaking 81–100% English in class.

Teacher 9 taught two English classes, of which the students in Group A completed the vocabulary as well as the grammar test. Group B took the grammar test only. As the results might reveal differences and similarities between each group but also between the grammar and vocabulary scores, the results of all three groups are reported individually in the following sections.

3.3.3.2.1 Group 9A Vocabulary

In group A, 18 students completed the vocabulary test at time one and 21 at time two. The interval between the test dates was 258 day or 8.5 months. Of the 18 students who completed the vocabulary at both test times, the scores of 8 decreased and 10 increased (*Figure 14*).

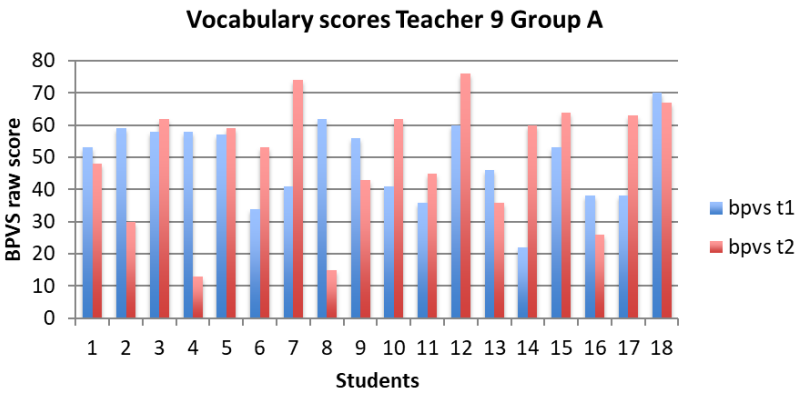

Figure 14 Vocabulary scores of Teacher 9's students at two times

There is considerable inter-individual as well as intra-individual variance, for example in students 2, 4, 7, 8, 10, and 14. The mean raw score for the vocabulary test of the paired participants at time one was M = 49 (SD = 12.11, min = 22, max = 70). At time two the mean raw score was M = 49.78 (SD = 18.60, min = 13, max = 76). A one-tailed paired samples t-test was not significant for the increase in vocabulary (t (17) = 0.14, p = .45).

3.3.3.2.2 Group 9A Grammar

At time one 18 students participated in the grammar test and 20 students at time two. The interval between the test dates was 222 days or 7.3 months. Of the 15 students who completed the tests at both test times, the grammar percentages of 11 students increased, 1 was unchanged, and 3 decreased (*Figure 15*).

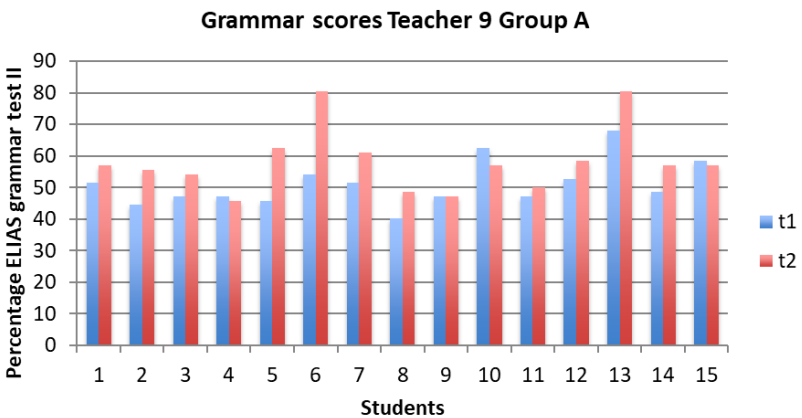

Figure 15 Grammar scores of Teacher 9's Group A at two times

The mean percentage for the paired student grammar test with group A at time one was M = 51.11 (SD = 7.01, min = 40.28, max = 68.06), at time two M = 58.15 (SD = 9.96, min = 45.83, max = 80.56). A one-tailed paired sample t-test was significant for the increase in grammar (t(14) = 1.89, p = .034).

3.3.3.2.3 Group 9B Grammar

In Teacher 9's group B, 16 students took part at time one and 20 at time two. The interval between the test dates was 222 days or 7.3 months. Of the 14 students who completed the tests at both test times, the grammar percentage of 7 students increased and 7 decreased (*Figure 16*).

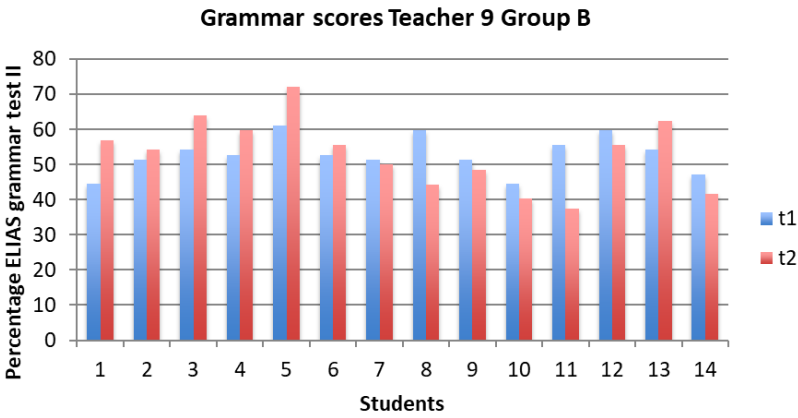

Figure 16 Grammar scores of Teacher 9's Group B at two times

The average grammar percentage of the paired participants at time one was M = 52.87 (SD = 5.02, *min* = 44.44, *max* = 61.11) and M = 53.08 (SD = 9.58, *min* = 37.5, *max* = 72.2) at time two. A one-tailed paired sample t-test was not significant for the increase in grammar ($t(13) = 0.28$, $p = .39$). This group had the highest mean grammar score of all the groups at time 1.

3.3.3.3 Classes Teacher 10 Results

Teacher 10 had two classes, of which group A was tested for grammar and group B for vocabulary. The teacher had not lived in an English-speaking country but held a degree in English. Teacher 10 reported an adaptive score 4.6 on the teacher questionnaire, a score 4 for speaking language proficiency, a score 5 for language proficiency for teaching, and reported speaking 61–80% English in class. It was the only teacher reporting the amount of English in class to be lower than 81–100%.

The school was in an urban area of a medium size town and had about 180 students. The school stated having children from 18 different heritage countries.

3.3.3.3.1 Group 10A Grammar

In Teacher 10's group A 15 students took the grammar test at time one and 20 at time two. The interval between the test dates was 218 days or 7.2 months. Of the 12 students who completed the tests at both test times, the percentages of 4 students increased and 7 decreased (*Figure 17*).

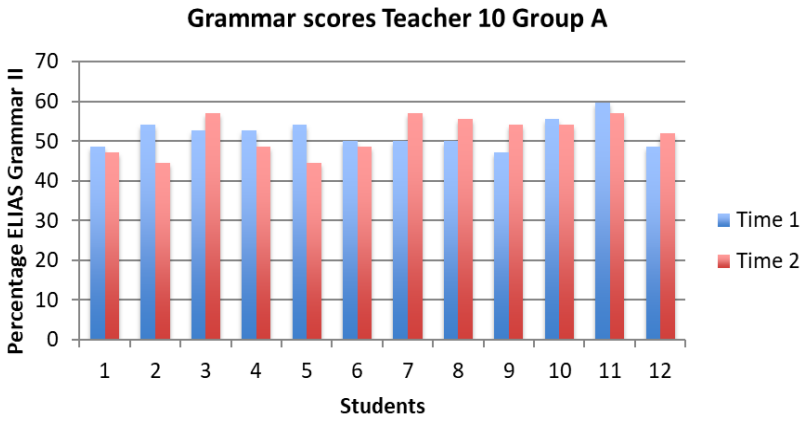

Figure 17 Grammar scores of Teacher 10's Group A at two times

The average grammar percentage of the paired participants at time one was M = 51.97 (SD = 3.42, min = 45.83, max = 59.72), at time two M = 51.67 (SD = 4.60, min = 44.44, max = 56.94). A paired samples t-test was not significant for the difference in grammar at both times ($t(11) = 0.349$, $p = .365$).

3.3.3.3.2 Group 10B Vocabulary

In Teacher 10's group B 17 students took part in the vocabulary test at time one, 18 at time two. The interval between the test dates was 222 days or 7.3 months. Of the 15 students who took part at time one as well as time two, the vocabulary raw score of 10 students increased, 2 stayed unchanged, and 3 decreased (*Figure 18*).

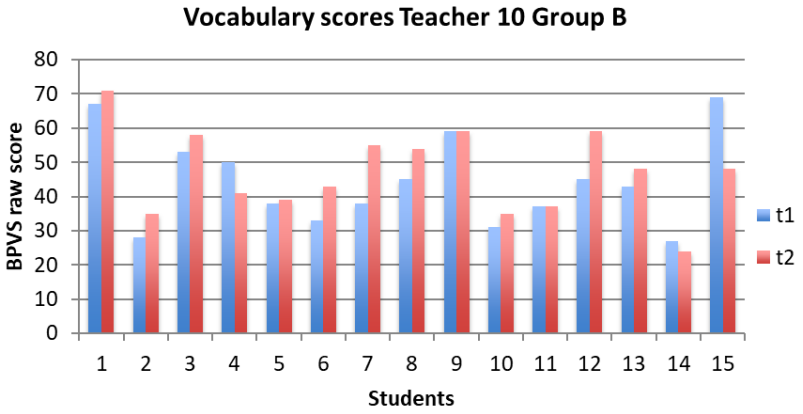

Figure 18 Vocabulary scores of Teacher 10's Group B at two times

At time one of the vocabulary test, the mean raw score of the paired participants was $M = 44.20$ ($SD = 12.79$, $min = 27$, $max = 69$) and $M = 47.07$ ($SD = 11.92$, $min = 24$, $max = 71$) at time two. A one-tailed paired samples t-test was not significant for the increase in vocabulary ($t(14) = 1.196$, $p = .126$).

The mean score of Teacher 10's group was between Teacher 9 and 11's group at both times, but this group gained the most of all groups in vocabulary from time 1 to time 2 (+2.87).

3.3.3.4 Classes Teacher 11 Results

Teacher 11 had not lived in an English-speaking country and did not hold a degree in English, but a degree in a different language. The school was in a rural area and had about 190 students. Teacher 11 reported an adaptive score 4.2 on the teacher questionnaire, a score 3.4 for speaking language proficiency, a score 5 for language proficiency for teaching, and reported speaking 81–100% English in class.

Teacher 11 taught two groups. Group A completed the vocabulary as well as the grammar test. Group B took the grammar test only.

3.3.3.4.1 Group 11A Vocabulary

In Teacher 11's group A 16 students participated in the BPVS test at time one, 15 students at time two. The interval between the test dates was 232 days and 7.6 months. Of the 15 students who completed the tests at both test times, the raw scores of 10 students increased from time one to time two, 5 decreased (*Figure 19*).

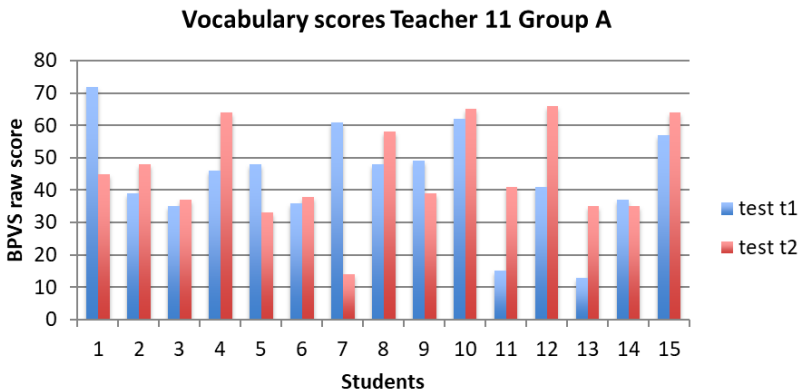

Figure 19 Vocabulary scores of Teacher 11's students Group A at two times

There is considerable inter-individual and intra-individual variance, for example in students 1, 7, 11, and 12. A high amount of variance in the vocabulary scores was already reported for teacher 9 (*Figure 14*) and to a lesser degree for teacher 10's group (*Figure 18*).

The mean raw score of at time one was M = 43.93 (SD = 16.12, min = 13, max = 72), at time two M = 45.47 (SD = 15.137, min = 14, max = 66). A one-tailed t-test was not significant for the vocabulary increase ($t(14)$ = 0.385, p = .186).

3.3.3.4.2 Group 11A Grammar

In Teacher 11's group A all 15 students took part in the grammar test at both times. The interval between the test dates was 229 days and 7.5 months. Of the 15 students who completed the tests at both test times, the percentage of 10 students increased from time one to time two, 5 decreased (*Figure 20*).

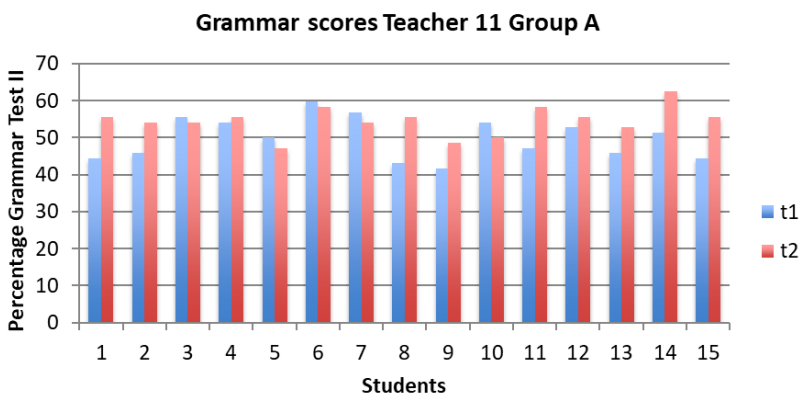

Figure 20 Grammar scores of Teacher 11's Group A at two times

The mean percentage of the paired participants at time one was M = 49.82 (SD = 5.41, min = 41.67, max = 59.72). At time two the mean increased to M = 54.54 (SD = 3.74, min = 47.22, max = 62.5). A one-tailed paired-sample t-test was not significant for the increase in grammar ($t(14)$= -0.86, p = .199).

3.3.3.4.3 Group 11B Grammar

Teacher 11's group B 16 students took part in the grammar test at time one, and 16 at time two. The interval between test dates was 232 days or 7.6 months. Of the 15 students who completed the tests at both test times, the percentage of 14 paired students increased from time one to time two, 1 decreased (*Figure 21*).

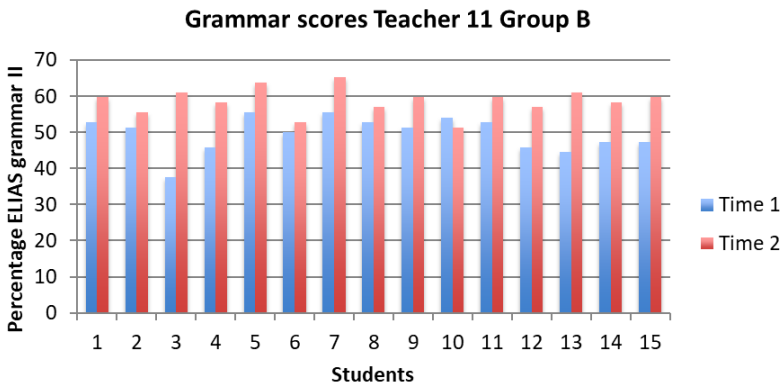

Figure 21 Grammar scores of Teacher 11's Group B at two times

The mean percentage of the paired samples at time one was $M = 49.8$ ($SD = 4.7$, $min = 37.5$, $max = 55.6$). At time two the mean increased to $M = 58.7$ ($SD = 3.57$, $min = 51.4$, $max = 65.3$). A one-tailed paired-sample t-test was significant for the gain in grammar ($t(14) = 3.96$, $p < .001$). This group had the highest mean score of all groups on grammar at time 2, but did not start out the highest at time 1.

3.3.3.5 Group Results

Figure 22 shows the mean vocabulary test results and their standard deviations of all test participants per teacher group at both times. The students of Teacher 1 did not sit the vocabulary test.

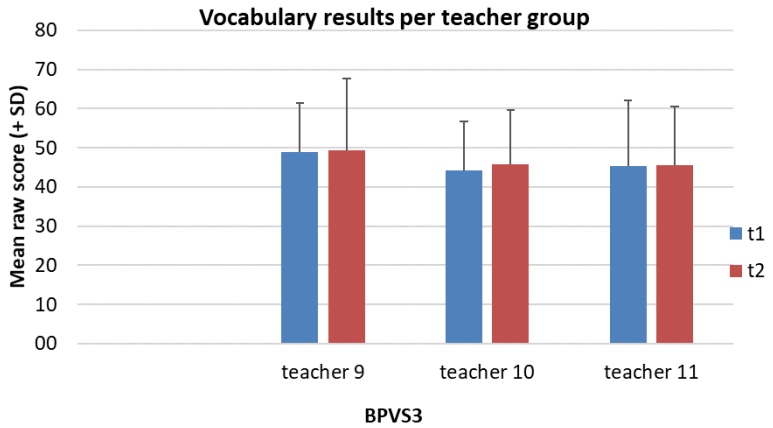

Figure 22 Vocabulary results per teacher group
Note. Teacher 9 t1 n = 18, t2 n = 21, Teacher 10 t1 n = 17, t2 n = 19, Teacher 11 t1 n = 16, t2 n = 15.

Figure 23 shows the mean grammar test results and their standard deviations of all test takers per teacher group at both times, based on the calculations already shown in Table 9.

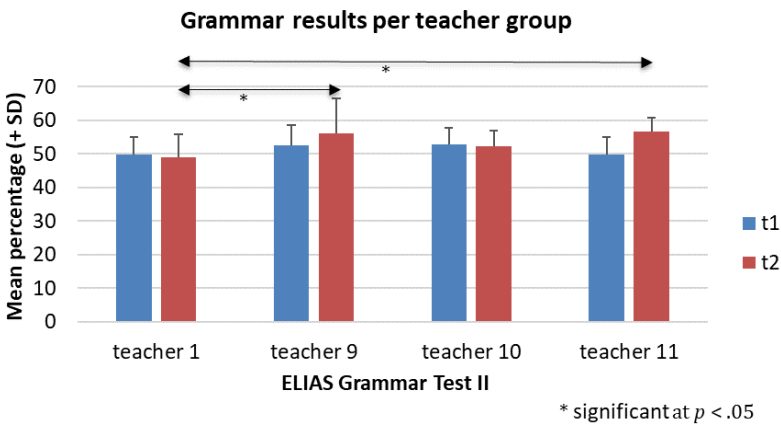

Figure 23 Grammar results per teacher group
Note. Teacher t1 n = 13, t2 n = 14. Teacher 9 t1 n = 34, t2 n = 40. Teacher 10 t1 n = 15, t2 n = 20. Teacher 11 t1 n = 32, t2 n = 30.

There was a mean grammar increase in the groups of teachers 9 and 11, whereas the means of the groups of Teacher 1 and 10 did not show any increase (*Figure 23*). The mean grammar scores at time two differed significantly between teachers 1 and 11 ($p = .016$) as well as between teacher 1 and 9 ($p = .022$). The significance is based on the results of the Tukey test of comparison in Table 11 explained there.

Teachers 9, 10, and 11 each taught two groups of students, Groups A and B respectively. To illustrate how each class of students scored on the tests, the paired means of the student groups per teacher at both test times is shown in *Figure 24* for vocabulary and in *Figure 25* for grammar.

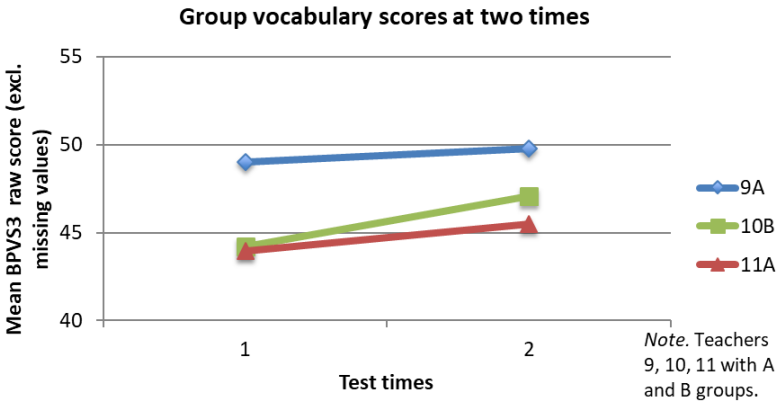

Figure 24 Student group vocabulary scores at two times
Note. Paired samples. Teacher 9A $n = 18$, Teacher 10B $n = 15$, Teacher 11A $n = 15$.

Figure 24 illustrates that student group 10B of Teacher 10 showed the most increase of the groups tested for vocabulary from time one to time two (+ 2.87), followed by Teacher 11's group (+ 1.54) and Teacher 9's group (+ 0.78), based on the calculations of the mean scores and the t-test results shown in the previous sections on the individual classes (sections 3.3.3.2.1, 3.3.3.3.2, 3.3.3.4.1). All three groups showed a positive tendency in their receptive vocabulary mean scores over the school year, while the reported t-tests did not reveal any significant differences between time one and two. Regarding the paired grammar means, the groups differed in their development. Group 11B showed the most increase from time one to time two, followed by group 9A and 11A. Groups 9B and 10A stayed about the same, as did group 1A (*Figure 25*).

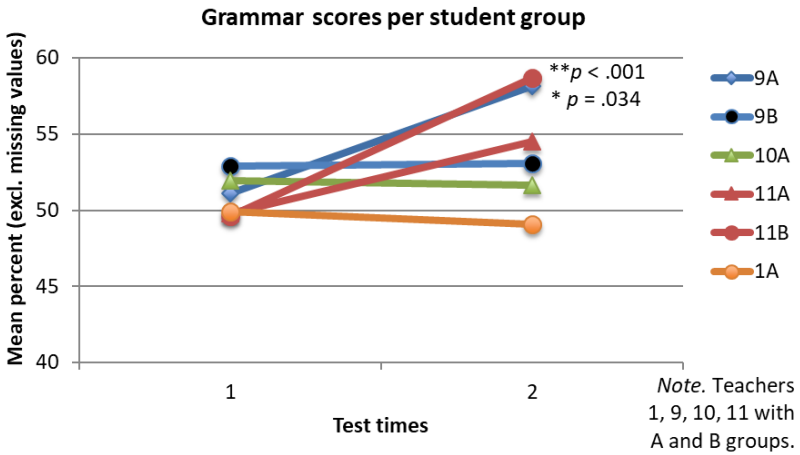

Figure 25 Group grammar scores at two test times
Note. Paired samples. Teacher 1A n = 13, Teacher 9A n = 15, Teacher 9B n = 14, Teacher 10A n = 12, Teacher 11A n = 15, Teacher 11B n = 15.

Teacher 9's Group 9A significantly improved from time one to time two by + 7.04%. Teacher 11's Group 11B showed a significant gain as well and improved by + 8.9%, based on the t-test results reported for all groups in the individual class results (sections 3.3.3.1 to 3.3.3.4.3).

For the following calculations, the students were sorted into groups with their respective teachers. A and B groups of the same teachers were combined in one group by teacher. *Figure 26* shows that Teacher 11's students gained + 6.90 percent in the mean grammar score between time one and two. Teacher 9's total student group gained + 3 percent on the mean grammar score. Teacher 10's group decreased -0.12 percent. Teacher 1's students decreased -0.96 percent.

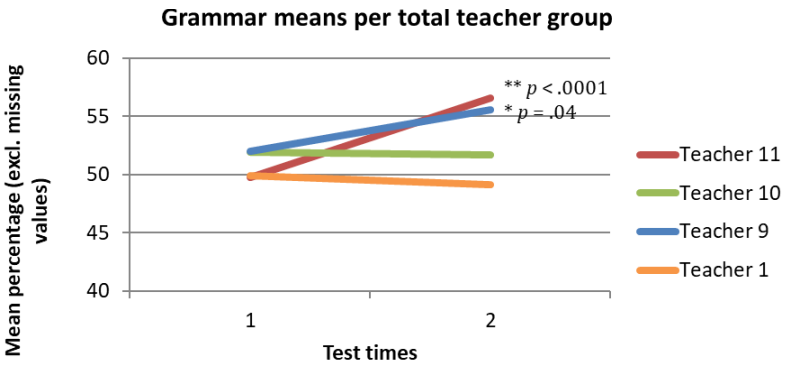

Figure 26 Grammar means per teacher group

Note. Paired samples. Teacher 1 $n = 13$, Teacher 9 $n = 29$, Teacher 10 $n = 12$, Teacher 11 $n = 30$.

To compute whether there was a significant difference between the teachers' student groups in vocabulary and grammar attainment at each time 1 and time 2, two one-way ANOVAs were conducted. For time 1, the one-way ANOVA with teacher as the independent variable and the students' vocabulary and grammar scores at time 1 as the dependent variables revealed no significant difference in the teacher effect between the groups on either test, vocabulary (F (2, 48) = .57, p = .569) nor grammar (F (3, 90) = 1.817, p = .15) (Appendix R).

The corresponding one-way ANOVA was run for time 2, with teacher as the independent variable and the students' vocabulary and grammar scores at time 2 as the dependent variables (descriptive statistics, see Appendix S). The results in Table 10 show there was no significant effect of teacher on the vocabulary scores at time two (F (2, 52) = 0.346, p = .709). Therefore, the students did not differ significantly in the vocabulary scores at time two between their respective teachers. A significant effect of teacher was found for the grammar scores at time two (F (3, 100) = 4.28, p = .007). The student groups differed significantly in their grammar scores at time two, when the students were sorted according to their respective teachers.

Table 10 *One-way ANOVA scores at time 2*

		Sum of Squares	*df*	Mean Square	*F*	Sig.
Raw score BPVS	Between Groups	177.278	2	88.639	.346	.709
	Within Groups	13309.558	52	255.953		
	Total	13486.836	54			
Percent ELIAS grammar test	Between Groups	747.221	3	249.074	4.280	.007
	Within Groups	5819.118	100	58.191		
	Total	6566.339	103			

Note. Vocabulary BPVS $N = 54$. Grammar $N = 103$

Because the one-way ANOVA for the grammar scores at time two showed a significant between-group difference, a post-hoc comparison Tukey HSD was computed to determine the nature of the differences between the teacher groups for the grammar test. Table 11 shows between which teachers' students significant differences occurred in the grammar test means (Appendix S).

Table 11 *Teacher group differences in grammar at time 2, Tukey comparisons*

Dependent Variable	(I) teacher	(J) teacher	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Percent ELIAS grammar test	1	9	-6.935*	2.369	.022	-13.124	-.745
		10	-3.046	2.658	.662	-9.991	3.890
		11	-7.513*	2.469	.016	-13.964	-1.062
	9	10	3.889	2.089	.251	-1.569	9.347
		11	-.5787	1.842	.989	-5.392	4.235
	10	11	-4.468	2.202	.184	-10.221	1.286

*$p < .05$.

The Tukey HSD results revealed how each teacher's group of students differed from one another. There was a significant difference (p = .022) in the grammar scores at time two between Teacher 1 and Teacher 9 with Teacher 1's group being significantly lower (-6.94).[26] Teacher 1's group also scored significantly lower (p = .016) than Teacher 11's group (-7.51). There were no statistically significant differences between Teacher 10's student group and any of the other teachers' groups and no statistically significant difference between Teacher 9's and Teacher 11's group. *Figure 27* illustrates the differences between the teachers' groups of students for the grammar scores at time two.

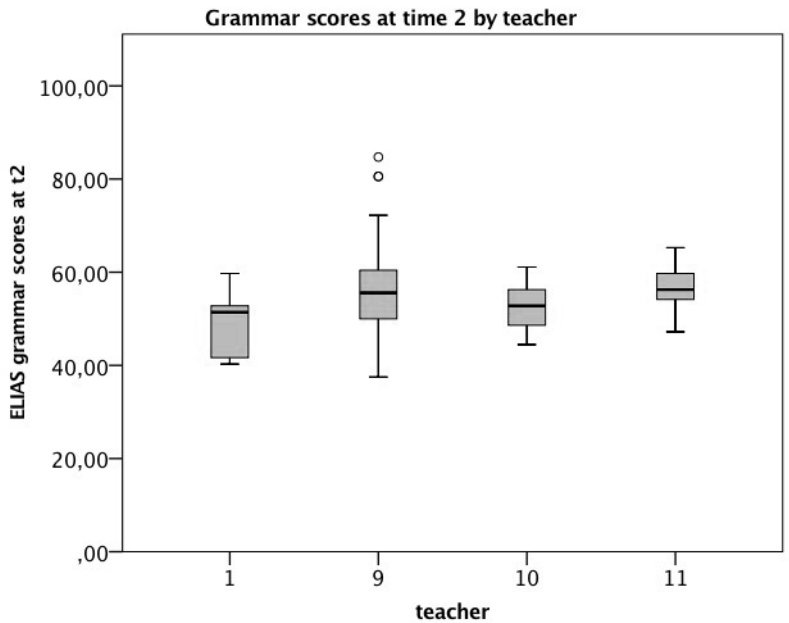

Figure 27 Boxplot grammar scores at time 2 by teacher

It was shown that significant differences in the test scores between the teachers' student groups were only found at time 2, not at time 1. As this result suggested that there was development from time 1 to time 2, two one-way ANOVAs were computed to find whether a significant teacher effect could also be found when the differences between the scores (t2-t1) were chosen as independent variables.

26 If the lower and upper confidence intervals (CI) do not pass through 0, the relation between the variables is statistically significant.

Two one-way ANOVAs each with vocabulary and grammar difference (t2-t1) as the respective independent variable and teacher as factor did not reveal statistically significant differences between the student groups in vocabulary development from time one and time two (p = .949). A significant difference was found again between the student groups in their grammar development (p = .007) (Table 12).

Table 12 *One-way ANOVA difference between t1 and t2*

		Sum of Squares	*df*	Mean Square	*F*	Sig.
Vocabulary difference t2-t1	Between Groups	37.814	2	18.907	.052	.949
	Within Groups	16742.594	46	363.969		
	Total	16780.408	48			
Grammar difference t2-t1	Between Groups	767.441	3	255.814	4.346	.007
	Within Groups	4885.940	83	58.867		
	Total	5653.381	86			

Note. Paired samples.

To sum up the main findings, there was a significant teacher effect with respect to the difference in the attainment of the grammar scores at the end of elementary school at time 2 as well as in the improvement of the grammar scores from time 1 to time 2. No statistically significant differences between the students sorted by teacher were found for vocabulary at either test time, nor in the difference score between the two test times. There were no statistically significant differences between the groups at time 1.

A post-hoc Tukey test was run to analyze which teachers' groups of students differed significantly in their grammar development from time one to two. It revealed significant differences between the groups of Teachers 1 and 11, and 10 and 11 (Table 13).

Table 13 *Grammar difference score (t2-t1) between teacher groups, Tukey comparisons*

Dependent Variable	(I) Teacher	(J) Teacher	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Grammar difference t2-t1	1	9	-3.956	2.523	.403	-10.572	2.660
		10	-.846	3.071	.993	-8.899	7.207
		11	-7.860*	2.548	.014	-14.539	-1.180
	9	10	3.111	2.597	.630	-3.699	9.920
		11	-3.903	1.950	.196	-9.016	1.209
	10	11	-7.014*	2.621	.044	-13.885	-.143

*p < .05.

As opposed to the difference between teacher groups at time two, the difference between Teacher 1's and 9's students was not significant for the grammar gain between the two test times (p = .403). In addition, there was a significant difference between the groups of Teachers 10 and 11 (p = .044), which did not show at time two. The difference between Teachers 1's and 11's students remained significant (p = .014). *Figure 28* illustrates the mean grammar difference between time one and time two by teacher.

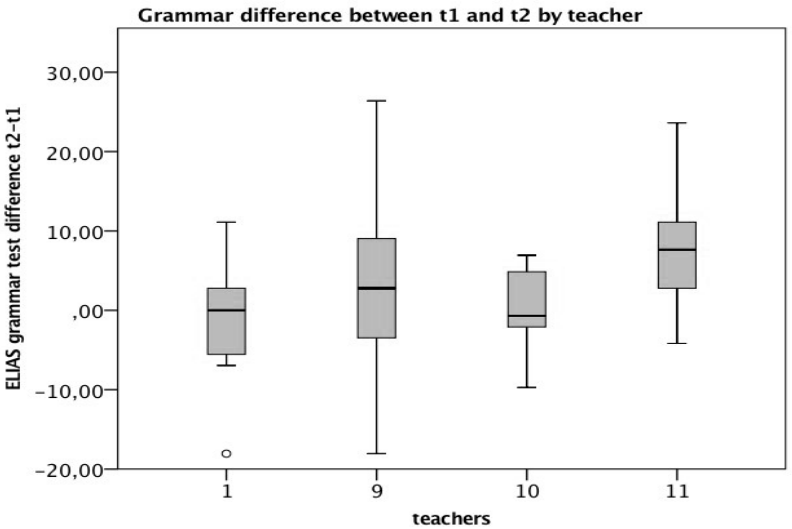

Figure 28 Boxplot grammar difference by teacher

The variance in the scores at the two test times as well as in the difference between time one and time two suggested there was a within-subjects effect. To confirm whether there was a systematic pattern of differences within individuals over the course of fourth grade, a repeated-measures ANOVA with time as a factor (t2-t1) was computed comparing the test scores of the students at both times. As was expected from the reported paired samples t-test results (Table 8), the repeated-measures ANOVA showed no significant time effect for the BVPS scores (F (1, 48) = .40, p = .529) (Appendix T) and a significant time effect for the grammar scores (F (1, 86) = 14.59, p < .001) (Appendix U).

3.3.3.6 Discussion of Study 2: Students' Receptive Vocabulary and Grammar

In the following, the findings of the test results are summarized and discussed. The research questions and hypotheses gearing Study 2 were the following: (RQ3) How do the students' receptive English grammar and vocabulary develop over their fourth year of elementary school? Hypothesis H3 predicted that the students' mean scores in vocabulary and grammar would increase over the course of the fourth grade. (RQ4) How do the student groups differ per teacher in their receptive English vocabulary or grammar attainment and development?

The findings showed that in total, the mean grammar percentage increased significantly from time one to time two. The mean vocabulary score increased non-significantly. This was revealed when all student participants were included as well as when only paired samples of those students who completed the respective test at both times were taken into the calculations. Only the mean grammar score increase over the school year was statistically significant. Hypothesis H3 was rejected for the development of the mean vocabulary score, which did not significantly increase between the two times, and supported for grammar, as it improved significantly. Further, there were significant differences between the student groups sorted by teacher.

Summarizing, the participants as a whole group, sampled from seven different classes at four different schools in different parts of Lower-Saxony, significantly improved their receptive grammar during their fourth year of elementary school as assessed in the ELIAS Grammar Test II. There was a non-significant positive tendency in the students' receptive vocabulary as tested in the BVPS3 vocabulary test in three classes taught at three different schools.

The results indicate that as a whole group, the students significantly improved in receptive grammar and tended to enhance their receptive vocabulary, albeit not statistically significantly, while being instructed in English for two hours a week during their fourth year of a regular elementary school program. As a number of studies have shown (Aukrust, 2007; Graham et al., 2017; Rohde, 2010; Steinlen et al., 2010; Unsworth et al., 2015; Weitz, 2015), more improvement may be expected if the number of hours of English instruction is higher.

Since the participants of the current studies had mainly been exposed to English in the classroom, their average L2 development may not trace back to extra-curricular language and speaker contact – based on the children's self-reports on English contact outside of class. In addition, a positive effect of out-of-class exposure to English such as subtitled TV in some countries (Enever, 2014; Huang et al., 2018; Kuppens, 2010; Lindgren & Muñoz, 2013), as explained in section 2.2.6, was not expected for the German learning context. Therefore, it is reasonable to conclude from the present study findings of the children's tests that is was the instruction of English as a second language at elementary school that may have supported the development of receptive L2 English in children, even at such a small amount of instruction as two lessons a week. Yet the children's inter-individual variability, the fact that some students did not improve their scores, and differences between groups of the same teacher suggest that an instruction effect may not be inevitable.

The results also indicated that on average, the selected children at the four investigated schools started out at comparable levels of receptive

vocabulary and grammar at the beginning of the study, regardless of the rural or suburban school location, different teachers, or varying school sizes in terms of the student body. There were no significant differences in the mean scores at the beginning of the study, when the students had already had prior exposure.

Having said that, the group results showed a more differentiated picture when the students were divided into their respective teacher groups. When the classes were separated according to their teachers and compared, there was some variance between the groups. Differences showed between the vocabulary and grammar attainment at time two as well as between the development from time one to time two.

At time one, the means of all groups were fairly close and the differences between them not significant: the mean vocabulary raw scores ranged from about 44 to 49, the mean grammar percentages from approximately 50 to 53 percent. In fact, three groups had approximately the same mean grammar scores at time one: Teacher 1's group and both of Teacher 11's group had mean scores of about 50 at time one for grammar.

The picture was changed at time two, when the groups showed unequal attainments. At the end of fourth grade, Teacher 9's students scored highest in vocabulary with a mean score of approximately 50, followed by Teacher 10 with about 46, and Teacher 11 with about 45. The means of the vocabulary scores had increased more in one group, Teacher 10's, than in the other two. The grammar attainment at time two showed a considerable range between all groups from about 49 to 59 mean percent. Two groups, Teacher 1's and Teacher 10's, did not improve their average grammar scores.

For the purpose of illustration, some comparison with studies using the same tests may put this study's participant scores into perspective. Compared to the receptive vocabulary scores reported in Couve de Murville et al. (2016), who examined four bilingual school groups using the BPVS, the mean scores of the participants of the present study were higher than the first and second graders: The first and second graders' mean scores in Couve de Murville et al.'s (2016) study were between 31 and 36 at time one and 34 and 44 at time two. Compared to the groups of third and fourth graders of an English immersive school, who had mean scores of 74 at time one and 75 at the end of fourth grade, the students of the current study scored much lower. Due to the smaller amount of target language exposure, which is one of the most outstanding differences between regular and immersive schooling, a lower score was to be expected for the students at regular elementary schools. Considering that the present students had been taught for only two lessons a week over a comparably short time interval from test time one to test time two, the positive tendency in receptive vocabulary between time one and time two is remarkable, even though statistically not significant.

As was expected for the grammar means as well, the students in the current study performed lower on the receptive grammar development than students studied in an immersion school context using the ELIAS grammar test (e.g. Wiegand, 2014). At time two, Teacher 11's total student group had the highest mean score of approximately 57 percent, followed by Teacher 9's total group with about 56 percent, Teacher 10's with about 52 percent, and Teacher 1's with about 49 percent. Comparing these results to a figure reported in a study by Buyl (2010), who used the ELIAS Grammar Test first edition, Buyl reported a mean grammar score of 58 percent after 18 months of immersive English primary school exposure. In the present study, Teacher 11's group as well as Teacher 9's group differed from Buyl's participants by one to two percent of the correct answers on the ELIAS Grammar Test II. However, the immersion students reached the respective scores at a younger age than the children in the present study. Further research at regular schools, ideally with similar test instruments, would help to obtain comparable scores.

In addition to variance in the mean scores between some of the teachers' groups, the individual children varied greatly in all groups, regardless of who their teachers were. A considerable amount of individual variance was reflected in the standard deviations as well as the students' individual scores at both times. Some students' scores increased, some decreased, others stayed unchanged. This was particularly the case with the vocabulary scores, but was also observed in the individual grammar scores. Variation between individual learners on receptive vocabulary and grammar has been reported in several other studies as well (Couve de Murville et al., 2016; Steinlen et al., 2010; Unsworth et al., 2015) (section 2.2.6).

The results obtained from the present study on receptive vocabulary and grammar are in line and lend support to previous research in that they also show a high degree of individual variation, while the total group mean may increase, as was reported for Couve de Murville et al. (2016) in section 2.2.6. Between-subject as well as within-subject variance is frequently attributed to individual learner differences manifest in components such as language aptitude, personality traits, emotional factors, learner styles and strategies (Dewaele, 2009; Dörnyei, 2014; Paradis, 2011; Skehan, 1991). A large amount of variance may indicate that other external or internal factors, which were not accounted for in the present study, were at play. Due to the focus of the study design to investigate teachers' language performance as a possible factor in students' L2 outcomes, specific factors affecting individual variation were not examined.

Another noticeable observation in the present study was that the significant differences between the groups' test scores were observed at time two and not at time one. The difference toward the end of fourth grade at

test time two can therefore not be attributed to large differences in receptive grammar and vocabulary already at the beginning of the study.

The analyses also revealed significant differences in the development of the mean scores between both times. Thus, not only differences between the test means of the teachers' groups at time two, but also the differences in the development between the test times indicated that the receptive L2 development during the fourth year of elementary school varied between the groups, depending on who taught the classes.

Input intensity has been mentioned as one of the factors often found to be an influencing factor in L2 language acquisition (e.g. Kersten, Schüle, et al., forthc. Lightbown, 2014; Maier et al., 2016; Muñoz, 2014; Rohde, 2010; Saito & Hanzawa, 2018; Steinlen et al., 2010; Unsworth, 2016a; Weitz et al., 2010). However, since the amount of English instruction time in the present study was the same for all students in each group, input intensity was expected to be similar for the students in their groups. Individual variation as well as differences between the groups occurred despite the same amount of instruction time. Except Teacher 10, who chose the second highest option on the questionnaire scale on in-class target language choice, the teachers reported speaking 81 to 100 percent English in class, which suggested that the amount of input in class was comparable between the groups. The self-reports did not show any significant differences between the individual teachers in their amount of target language choice in the classroom. However, the reliability of the self-reported answers particularly with respect to the amount of target language in the classroom needs to be considered as limited (see section 3.2.5.3).

A more detailed analysis was carried out on the teachers' linguistic performance, Part I of the synthesis Study 3, and the possible regularities between the teachers' CAF performance and the students' test findings, Part II of Study 3 (section 3.4.2). The synthesis Study 3 shows whether any relationships between the teachers' language performance and the children's test findings were found.

Looking at the difference between the development of receptive grammar and vocabulary, the present results suggest that, as a whole group, the children's receptive grammar developed more over the course of the school year than their receptive vocabulary. Only the gain in grammar between time one and time two was statistically significant. The differences between the development in receptive vocabulary and grammar may be attributed to the respective test properties of the ELIAS Grammar Test II and the BPVS3. As discussed in section 3.3.1, each test instrument has limitations that might not have been accounted for. The test instruments may also assess development differently for grammar than for vocabulary. Although the BPVS3 is also used for speakers of English as a second language, it was developed for English speakers in the UK. The ELIAS Grammar Test

on the other hand has been in use primarily in European preschool and elementary school settings. Ceiling effects on the BPVS are highly unlikely for early second language learners, as the prompts are numerous and include more infrequent words as the sets go up. However, the test is limited in its transferability to instructed foreign language acquisition settings, as was explained in section 3.3.1.2.1. This limitation has been commented on for example in Couve de Murville et al. (2016) as well.

The ELIAS Grammar Test II, on the other hand, is more limited in the amount of phenomena prompted. The odds of chance answers are higher as well, since the ELIAS test used three-pictured templates (33.3% chance of correct answer) while the BPVS has four pictures in each template (25% chance). In addition, the children's grammar results in the current study are based on a larger participant number than the group of vocabulary test takers, which may result in statistical effects showing more in the larger grammar group (Wang, Watts, Anderson, & Little, 2013).

As was suggested in section 2.2.6, vocabulary and grammar are expected to develop alongside in second language acquisition. Yet, like in the current findings, a difference between the rate of receptive grammar development and vocabulary was also found in Weitz et al. (2010). They have reported an effect of the input, as measured in an input quality observation scheme, on the development of the receptive grammar in their preschool participants, but not on their receptive vocabulary. Another study on the preschool children's L2 development revealed the intensity of L2 contact to predict change in receptive vocabulary, but not in grammar (Kersten, Schüle, et al., forthc.). Whether the observed difference between the gains of receptive grammar and vocabulary of the current study can be generalized to a greater population of fourth-graders, may therefore be the subject of further future research.

As was argued in section 2.2.6, however, comparability between immersion and bilingual preschool and regular elementary school children is limited due to greatly varying learning conditions. More studies incorporating the tests in use would be needed to argue a clear case for or against a systematic difference in the tested vocabulary and grammar development and, if there was one to be found, examine why there might be a consistent difference.

To sum up, the test results of the BPVS3 and the ELIAS Grammar Test II showed that the development of those two receptive target language areas over the fourth year at a regular elementary school tended to progress in the entire group of participants, varied individually, and showed differing results between the teachers' groups, not statistically significantly for receptive vocabulary and statistically significantly for receptive grammar.

The test results of the students in the present study revealed at least two levels that are relevant to the analyses following in the synthesis Study 3

below: First, there was a positive development in both receptive areas of vocabulary and grammar – not statistically significant for vocabulary and statistically significant for grammar – in terms of the means of the entire group of participants. Second, significant differences were found between the groups when they were analyzed separately according to the teachers who taught them.

The teachers' linguistic performance is analyzed in more detail in Part I of the following synthesis study. Their CAF performance is operationalized to reveal specific relationships between the dimensions and to obtain CAF scores that are used in the subsequent final part of the synthesis study. The connection between the teachers' L2 language performance and the students' receptive vocabulary and grammar is drawn in Part II of the synthesis Study 3.

3.4 Study 3: Synthesis of Study 1 and Study 2

This final empirical part focuses on a synthesis of the previous two studies. It first provides insights into specific features of the teachers' spoken language in terms of complexity, accuracy, and fluency. Second, it analyzes their possible effects on their students' second language acquisition.

The synthesis of the two studies aims to add a novel angle to connecting teachers' linguistic performance and foreign language development, as it incorporates the three CAF dimensions of performance as potential indicators of linguistic proficiency of the target language providers – the teachers in the study – at selected elementary schools in Germany. The results could have implications for foreign language teacher education regarding their own language training in terms of their complexity, accuracy, and fluency, their language development as well as their classroom language use as teachers. The synthesis study will look at the following research questions and hypotheses:

(RQ5) How can the CAF dimensions be transformed into a scale that can be used for further analyses?

(RQ6) How do complexity, accuracy, and fluency in the teacher's L2 performance relate to one another? As a correlation between the CAF dimension has been found in studies examining individual CAF performance, the following hypothesis was predicted: (H6) All three CAF dimensions correlate.

(RQ7) How does the teachers' L2 English performance, as measured in complexity, accuracy, and fluency, relate to their students' L2 receptive vocabulary and grammar development? Studies have not yet examined relationships between teachers' L2 performance measured within the CAF framework and their students' L2 development. There are no comparable

studies to this date that are based on similar statistical procedures in the analysis of the individual CAF dimensions. However, performance was discussed as being a part of over-all language proficiency, which has been argued to have an effect on the students' L2 development. In addition, hypothesis H6 suggested a correlation between the CAF dimensions. Therefore, the following was predicted: (H7) There is a positive relationship between the teachers' CAF performance as well as each of the CAF dimensions and the students' receptive grammar and vocabulary development.

(RQ8) If there is a relationship between teachers' L2 performance and children's foreign language acquisition, is there an additional effect by the classroom L2 use as rated by the teachers? Theoretical considerations on teacher-talk characteristics have suggested that simplified input language may benefit children's L2 development. Therefore, the following hypothesis was tested: (H8) The teachers' adapted L2 use in the classroom moderates a possible CAF effect on the children's receptive grammar and vocabulary development.

Two main parts were necessary in order to first provide a more detailed and more reliable analysis of the teachers' linguistic CAF performance values calculated in Study 1, and second, to be able to merge results of the interview Study 1 data on the teachers' performance with those of Study 2, the students' data on their receptive grammar and vocabulary results. Those two synthesis approaches are presented in Part I and Part II of this chapter.

Part I answers research questions RQ5 and RQ6 including hypothesis H6. The objective of Part I was to reduce the indices applied to measure the dimensions complexity, accuracy, and fluency of Study 1 and to obtain a single score for each dimension. The results of Study 1 in section 3.2.3 have already highlighted selected measures in the CAF dimensions and the teachers' scores. As discussed, the results could not give a comprehensive picture of each of the CAF dimensions because they focused on selected individual measures only. The theoretical section 2.1 has shown that research in the field of CAF is not conclusive. In particular, studies are not decided on the question whether or not there has to be a trade-off effect between complexity, accuracy, and fluency in second language production. Partly responsible for varying findings are different measures as well as different study designs employed in the studies.

Part I of this chapter adds a novel angle to operationalizing language performance in terms of CAF by including all measures underlying each of the dimensions. A composite score for each CAF dimension based on all performance measures aimed to include the measures according to the load they contributed to the respective dimension. Instead of excluding certain measures, or using only selected measures, all measures were included to analyze how relevant each measure was to its particular CAF dimension

and at what degree. Therefore, a Principal Component Analysis (PCA) was computed, described in more detail in section 3.4.1 below.

A Principal Component Analysis explains the variance-covariance structure of a set of variables through linear combinations. PCA is a dimensionality reduction technique, which "is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables" (Jolliffe, 2002, p. 1). Put differently, in a PCA, the number of the otherwise too large set of CAF variables is condensed into new sets of composite scores for each dimension, which can then be used in further data analyses, for example in the regression analyses shown in Part II of this synthesis study.

Part II answers research questions RQ7 and RQ8 and tests the respective hypotheses H7 and H8. The focus in Part II was to relate the teachers' language CAF performance to the students' language skills as measured in the receptive tests in Study 2. The composite scores gained from the PCA for each of the dimensions were, in a final step, used in the synthesis analysis between the CAF composite scores and the students' test results. Thus, in Part II of the synthesis, the principal component scores of the teachers' three CAF dimensions were examined in relation to the students' test scores, using regression analyses.

3.4.1 Study 3 Part I: Data Analysis of Principal Components and CAF Relations

For the data of the Study 1 teacher interviews the relations between the measured outcomes of each of the three CAF dimensions were analyzed, using Spearman's rho test of correlations. Due to the abundance of CAF measures in the teachers' interview Study 1, multicollinearity between the measures needed to be analyzed to then gain singularity of each CAF dimension. The lack of such an analysis in CAF studies has been criticized by Norris and Ortega (2009). In a PCA, the inter-correlated quantitative variables of the CAF analyses can be integrated in component scores for each CAF dimension.

PCA is a technique that can operate qualitative variables as a correspondence analysis and deal with heterogeneous sets of variables as a multiple factors analysis (Abdi & Williams, 2010).[27] By means of features extraction, the variables are assigned principal components of each

27 For a discussion of the main differences between a PCA and a factors analysis, see Little (2013).

dimension while including all the variables' loads. The main idea of a PCA is "to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set" (Jolliffe, 2002, p. 1).

By following this approach, the data analysis revealed the associations between the measures underpinning each dimension as well as the results of the Principal Component Analysis in each dimension. SPSS was used to run the analyses.

3.4.1.1 Results of Principal Components and CAF Relations

The following sections first show the results of the Principal Component Analysis for all CAF measures in the performance model. Second, the results of each individual dimension of complexity, accuracy, and fluency including its subdimensions breakdown, speed, and repair fluency are presented.

3.4.1.1.1 Performance

The initial principal components extraction including all the CAF variables produced four components (Appendix CC). Only components with eigenvalues greater than 1 were kept in the matrix. The eigenvalues pertain to the total variance explained by each component. The first principal component accounts for the largest amount of the variance (Wang et al., 2013). The variance load of the first principal component was 56.9% (Appendix CC). Table 14 shows the variables sorted by loading size onto the first principal component of performance from loading most to least. Because the component values are correlations, possible values range from -1 to +1. The farthest from zero a coefficient in either direction is, the largest the correlation with the component.

Table 14 *All measures, PCA component 1 loadings on CAF performance*

Dimension	Measure	Component 1
Accuracy	Error-free clauses per AS-unit	.938
Syntactic complexity	Words per AS-unit	.910
Accuracy	Percentage error-free clauses	.888
Breakdown fluency	Ratio pause total duration	-.862

Dimension	Measure	Component 1
Syntactic complexity	Subordinate clauses per recorded minute	.854
Speed fluency	Speechrate syllables per total duration	.849
Syntactic complexity	Ratio subordinate clauses to AS-units	.830
Speed fluency	Mean length of runs	.815
Breakdown fluency	Mean pause duration (sec)	-.783
Syntactic complexity	Clauses per AS-unit	.717
Lexical complexity	Lexical diversity D vocd	.598
Speed fluency	Articulation rate syllables per phonation time	.438
Repair fluency	Repairs per minute total dur.	-.430
Repair fluency	Repetitions per minute total dur.	-.245

The component matrix reveals that the first principal component was strongly correlated with eleven of the performance variables if a correlation above .5 or below -.5 is used as the cut-off point (Table 14). This suggests that those eleven variables varied together. The articulation rate as well as repairs and repetitions loaded much less strongly onto the first component.

Negative values were found for some fluency measures because the indices were defined negatively. Repairs per minute and repetitions per minute were measures of repair fluency, and mean pause duration as well as the ratio of pauses measured breakdown fluency. Therefore, negative values on those measures indicate high fluency. A factor score was created based on the loadings of the first principal component. This new variable score was named *performance*.

The second principal component was only correlated with two variables – repetitions per minute total duration and lexical diversity D vocd (Appendix CC). These two are expected to vary together in opposite directions, as the former's value is positive and the latter's is negative. The variance load of the second component was 14.1%. The third principal component was correlated only with repairs per minute total duration. The variance load was 9%. The fourth principal component was correlated only with articulation rate as syllables per phonation time and a variance load of 8.6%.

3.4.1.1.2 Complexity

Results of the two-tailed Spearman's rho correlation tests showed that there were several significant correlations among the syntactic measures. The ratio of subordinate clauses per minute correlated significantly with words per AS-unit (r_s (9) = .882**, p < 001). Each of these variables also correlated with two other syntactic measures – the ratio of subordinate clauses to AS-units and the ratio of clauses per AS-unit (Table 15). The vocd values indicating lexical diversity did not show any significant correlation with any of the syntactic measures.

Table 15 *Correlations of complexity measures*

		Ratio subordinate clauses to AS-units	Subordinate clauses per recorded minute	Lexical diversity D vocd	Words per AS-unit
Subordinate clauses per recorded minute	r_s	.845**			
	p	.001			
Lexical diversity D vocd	r_s	.118	.355		
	p	.729	.285		
Words per AS-unit	r_s	.700*	.882**	.573	
	p	.016	.000	.066	
Clauses per AS-unit	r_s	.473	.718*	.227	.764**
	p	.142	.013	.502	.006

N = 11
*p < .05. **p < .01. (2-tailed).

Table 16 shows the loadings of each complexity measure on the first principal component. For complexity, the first principal component showed high loadings of words per AS-unit, subordinate clauses per recorded minute and ratio of subordinate clauses to AS-unit. Clauses per AS-unit followed with a high loading of .76 and lexical diversity D loaded the least with .45 (Appendix BB).

Table 16 *Complexity loadings on component 1*

	Component 1
Ratio subordinate clauses to AS-units	.907
Subordinate clauses per recorded minute	.913
Lexical diversity D vocd	.449
Words per AS-unit	.949
Clauses per AS-unit	.757

Note. Extraction Method: Principal Component Analysis.

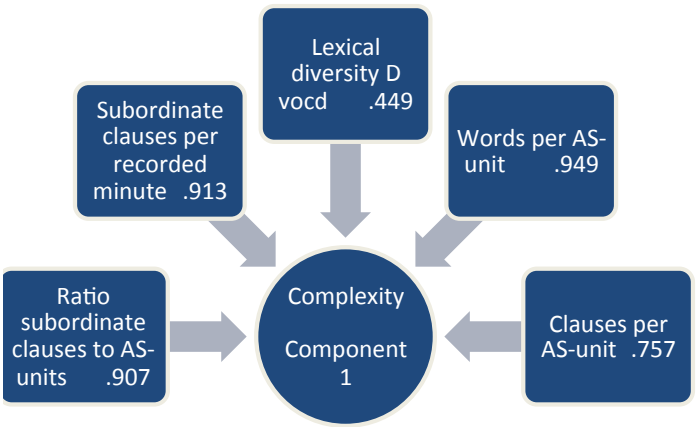

Figure 29 Complexity principal component 1

Figure 29 illustrates the loadings of each of the measure on the first component of complexity. The measured values in the first complexity component accounted for 66.63% of the variance in complexity performance. A composite factor score was computed based on the first principal component loadings. The new score was named *complexity comp.*

3.4.1.1.3 Accuracy

The two measures of accuracy, percentage of error-free clauses and the ratio of error-free clauses to AS-unit, correlated significantly ($r_s(9) = .907^{**}, p < 001$). Since the two measures of accuracy were significantly correlated with one another, both loaded equally onto the accuracy dimension (Appendix AA). *Figure 30* illustrates the loadings of the two accuracy measures.

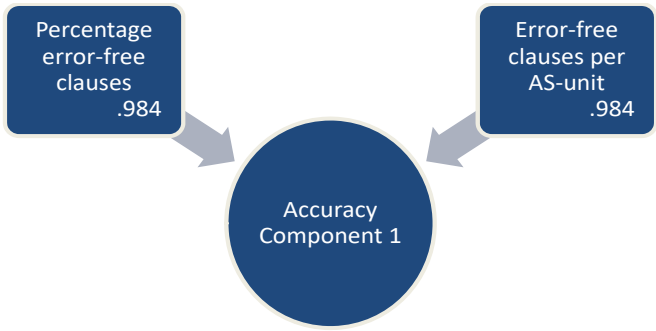

Figure 30 Accuracy principal component 1

The two accuracy measures account for 96.8% of the variance in the accuracy performance. A composite factor score was computed based on the first principal component and was named *accuracy comp.*

3.4.1.1.4 Fluency

The fluency measures were broken down into the subdimensions speed fluency, breakdown fluency, and repair fluency. Correlations between the measures of each fluency subdimension were analyzed using two-tailed Spearman tests.

Table 17 *Correlations speed fluency*

		Speech rate (syllables per total duration)	Pruned speech rate	Articulation rate (syllables per phonation time)
Pruned speech rate	r_s	.736**		
	p	.010		
Articulation rate (syllables per phonation time)	r_s	.809**	.355	
	p	.003	.285	
Pruned articulation rate per phonation time	r_s	.700*	.955**	.400
	p	.016	.000	.223

$N = 11$
*p < .05. **p < .01. (2-tailed).

Speech rate correlated significantly with all the other speed fluency values (Table 17). In addition, the pruned speech rate correlated significantly with the pruned articulation rate, but not with the articulation rate. The articulation rate correlated significantly only with speech rate.

Table 18 *Speed fluency loadings on component 1*

	Component 1
Speech rate syl./total dur.	.947
Pruned speech rate	.898
Articulation rate syl./phon. time	.682
Pruned art. phon. time	.918

Extraction Method: Principal Component Analysis.

Table 18 shows the loadings of the measures for speed fluency. The first principal component correlates significantly with all four measures and most strongly with speech rate, pruned articulation rate, and pruned speech rate. Articulation rate was significant, but less strongly (.68). The first component loaded 75.3% of the variance on the speed fluency dimension (Appendix X). *Figure 31* illustrates the loadings of the speed fluency measures. A composite factor score was computed based on the first component and named *speed fluency comp.*

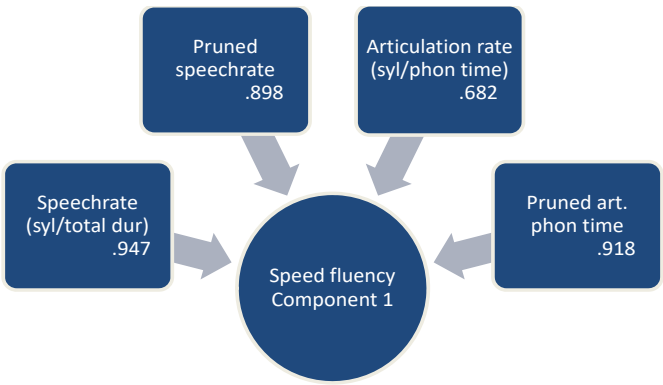

Figure 31 Speed fluency principal component 1

Breakdown fluency is another subdimension of fluency. Its measures are based on pauses.

Table 19 *Correlations breakdown fluency measures*

		Mean pause duration (sec)	Ratio pause total duration	Filled pauses per minute total duration
Ratio pause total duration	r_s	.773**		
	p	.005		
Filled pauses per minute total duration	r_s	.173	.227	
	p	.612	.502	
Mean length of runs	r_s	-.509	-.845**	-.391
	p	.110	.001	.235

$N = 11$
*p < .05. **p < .01. (2-tailed).

Of the breakdown fluency measures, the mean pause duration and the ratio of pauses per duration correlated significantly (Table 19). In addition, there is a significant negative correlation between mean length of runs and the ratio of pauses to total duration. Filled pauses per minute did not show any significant correlation with any of the other values measuring breakdown fluency.

Table 20 *Breakdown fluency loadings on component 1*

	Component 1
Mean pause duration (sec)	.850
Ratio pause total dur.	.957
Filled pauses per minute total dur.	.334
Mean length of runs	-.851

Extraction Method: Principal Component Analysis.

Table 20 shows the loadings of the measures for breakdown fluency. The first principal component increases with the mean pause duration, the ratio of pauses to total duration, and the mean length of runs, but not with

filled pauses per minute. This component correlates most strongly with the ratio of pauses to total duration, and strongly with mean pause duration and mean length of runs. The first principal component loads 61.81% of the variance on the breakdown fluency dimension (Appendix Y). *Figure 32* presents the loadings of breakdown fluency. A composite factor score was calculated based on the first principal component and named *breakdown fluency comp.*

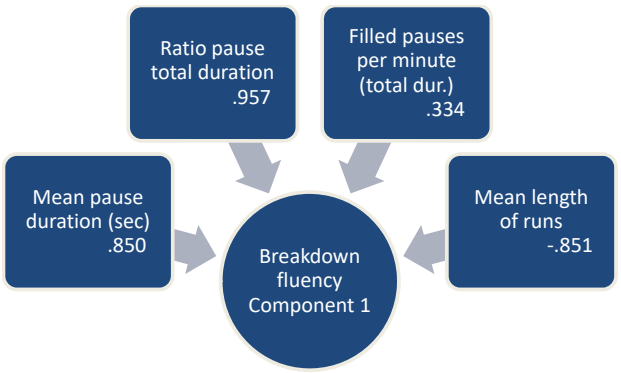

Figure 32 Breakdown fluency principal component 1

Finally, the two measures for repair fluency did not correlate significantly with one another ($r_s = 0.236$, $p = .484$). Table 21 shows that the first principal component correlates significantly with both repairs and repetitions per minute total duration.

Table 21 *Repair fluency loadings on component 1*

	Component 1
Repairs per minute total dur.	.804
Repetitions per minute total dur.	.804

Extraction Method: Principal Component Analysis.

The first component loaded 64.57% on the repair fluency dimension (Appendix Z). *Figure 33* presents the loadings of the two repair fluency measures. The composite factor score was calculated based on the first principal component and named *repair fluency comp.*

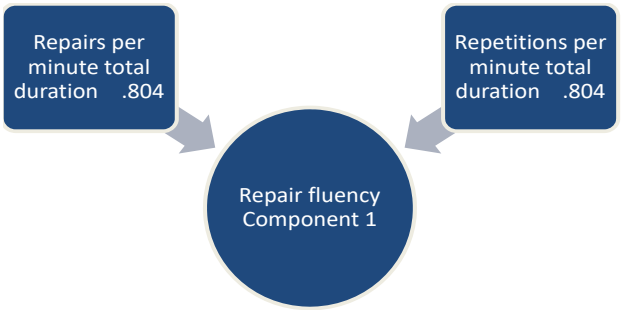

Figure 33 Repair fluency principal component 1

To sum up, the Principal Component Analysis revealed high loadings of most of the measures. The measures were included in a new composite score for each teacher on each dimension, based on the measure loadings on the first principal component of each dimension and subdimension. Factor scores reduced all the measured CAF values to a score that could be used in subsequent calculations, namely the relationships between the CAF dimensions on the one hand, and the relationships between the teachers' performance and their students' results on the other hand.

Research question RQ5 asked how the CAF dimensions could be transformed into scores. Principal Component Analysis proved to enhance the analysis of complexity, accuracy, and fluency in a way that includes the relevant measures for each dimension to form a composite score that can be used in further analyses. After calculating the principal components to each CAF dimension and creating new composite scores, it was possible to compute the relationships between each of the CAF dimensions. The relationships are analyzed in the following section.

3.4.1.1.5 Relationships Between CAF

As mentioned in section 2.1, studies on the relationships between complexity, accuracy, and fluency are not conclusive. One of the prominent questions has remained as to whether L2 speakers can perform similarly on all three dimensions simultaneously, or whether the CAF dimensions inevitably trade off. This section examines the relationships between the CAF dimensions based on the teacher interview results of Study 1. The following also illustrates how all eleven teacher participants scored and where those four teachers ranged whose students took part in Study 2 on the students' receptive vocabulary and grammar.

Each teacher was assigned a composite factor score per CAF dimension based on the Principal Component Analysis results as reported in the

previous sections. The teachers' complexity, accuracy, and fluency indices were now based on the principal components for each dimension and subsumed in a factor score, which is a composite score for each dimension integrating each underlying measure based on the magnitude of its load. Each teacher had a score for *complexity comp, accuracy comp, speedfluency comp, breakdown fluency comp,* and *repair fluency comp* (Appendix DD). In addition, lexical diversity vocd was included as the original values to reveal its relationships, because vocd did not correlate with the other complexity measures. Lexical diversity was based on the vocd score.

Table 22 shows the Spearman's rho correlation analysis results between the composites scores of complexity, accuracy, speed fluency, and breakdown fluency. Repair fluency comp was not significantly correlated with complexity and accuracy, nor was lexical diversity.

Table 22 *Correlations between CAF composites*

		Speed fluency comp	Breakdown fluency comp	Repair fluency comp	Accuracy comp	Complexity comp
Breakdown fluency comp	r_s	-.864**				
	p	.001				
Repair fluency comp	r_s	-.455	.418			
	p	.160	.201			
Accuracy comp	r_s	.764**	-.709*	-.345		
	p	.006	.015	.298		
Complexity comp	r_s	.845**	-.818**	-.491	.936**	
	p	.001	.002	.125	.000	
Lexical diversity D vocd	r_s	.673*	-.700*	-.664*	.336	.445
	p	.023	.016	.026	.312	.170

N = 11
*p < .05. **p < .01. (2-tailed).

Simple linear regression analyses calculated whether the teachers' CAF composites significantly predicted one another. The following graphs illustrate the relationships between each of the significant CAF dimensions based on the PCA results. The graphs also show where the individual

teachers ranged and among those, indicate the teachers whose students were tested in Study 2.

The more scattered from the regression line the points in the graph are, the less variance between the variables is accounted for. R^2 represents the effect size of the correlation and shows how much of the variance in the variables is accounted for. The strength of the effect sizes is not a fixed value but rather interpreted depending on the research field (Larson-Hall, 2016, p. 208f.). The closer R^2 is to 0, the less variance between the variables can be accounted for by the variables. The closer R^2 is to 1, the more variance in one variable is accounted for by the other variable.

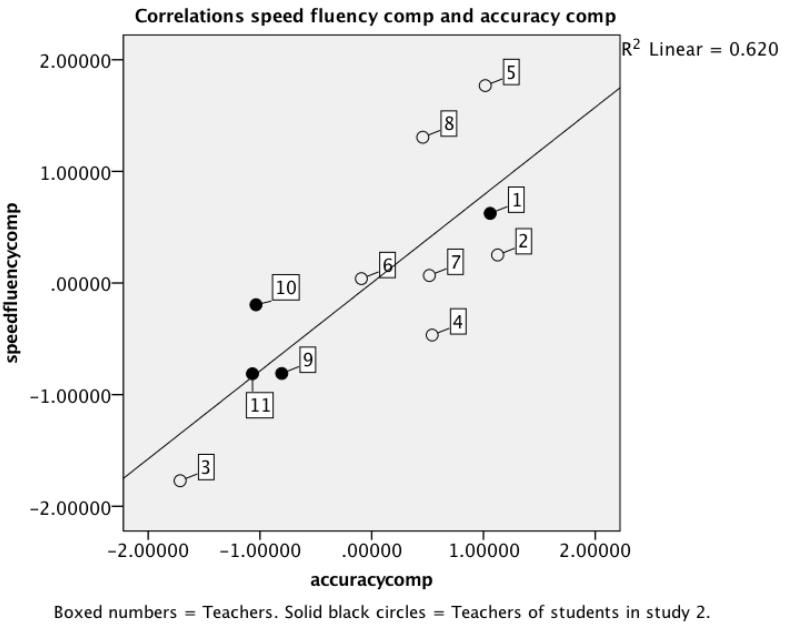

Figure 34 Scatterplot of teachers' speed fluency and accuracy

Figure 34 illustrates a significant positive linear regression between speed fluency and accuracy ($F(1, 9) = 14.687$, $p = .004$) with an R^2 of 0.62. Thus, 62% of the variance in speed fluency was explained by the variability in accuracy. When accuracy increased, speed fluency increased and vice versa. Teachers 1, 9, 10, and 11, who taught the students tested in Study 2, individually showed similar performances on each dimension respectively (Appendix EE).

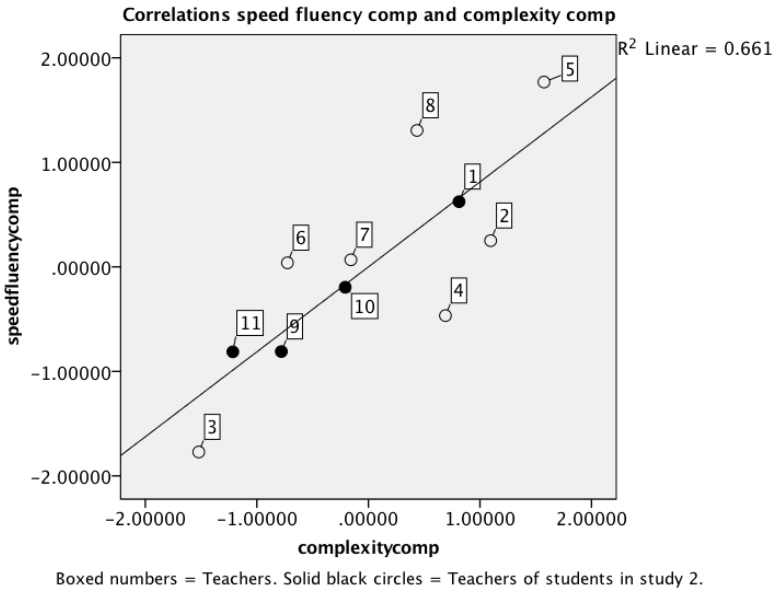

Figure 35 Scatterplot of teachers' speed fluency and complexity

There was a significant positive linear regression between speed fluency and complexity as well ($F(1, 9) = 17.516$, $p = .002$) with an R^2 of 0.661 (*Figure 35*). 66% of the variance in speed fluency was explained by the variability in complexity. When complexity increased, speed fluency increased and vice versa.

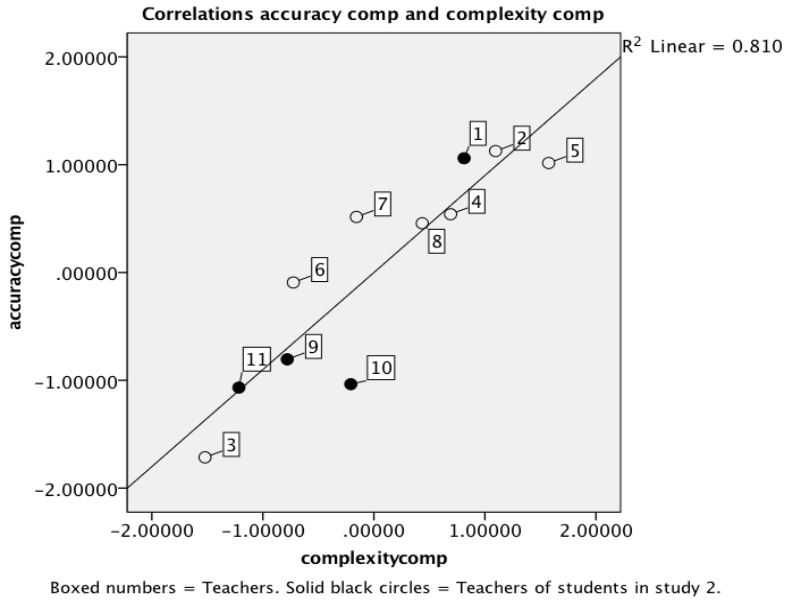

Figure 36 Scatterplot of teachers' accuracy and complexity

There was a significant positive linear regression between accuracy and complexity ($F(1, 9) = 38.444$, $p < .001$) with an R^2 of 0.810 (*Figure 36*). 81% of the variability in accuracy was explained by the variability in complexity. Accuracy increased when complexity increased and vice versa.

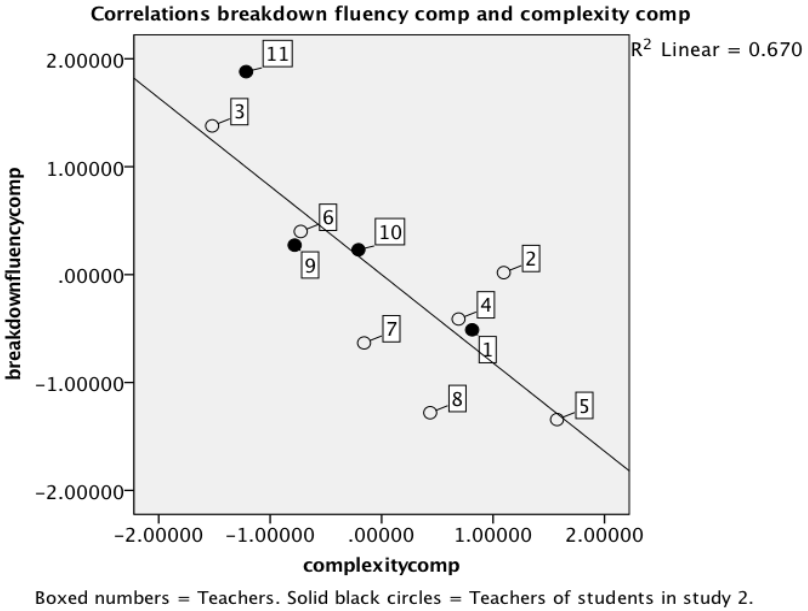

Figure 37 Scatterplot of teachers' breakdown fluency and complexity

There was a significant negative linear regression between breakdown fluency and complexity ($F(1, 9) = 18.282$, $p = .002$) with an R^2 of 0.670 (*Figure 37*). 67% of the variability in breakdown fluency was explained by the variability in complexity. When complexity increased, breakdown fluency decreased and vice versa.

The negative regression line is due to the negative measure of breakdown fluency. A high value on breakdown fluency indicates less fluent speech. Thus, the regression lines of breakdown fluency run in negative directions. High breakdown fluency indicates less fluent speech.

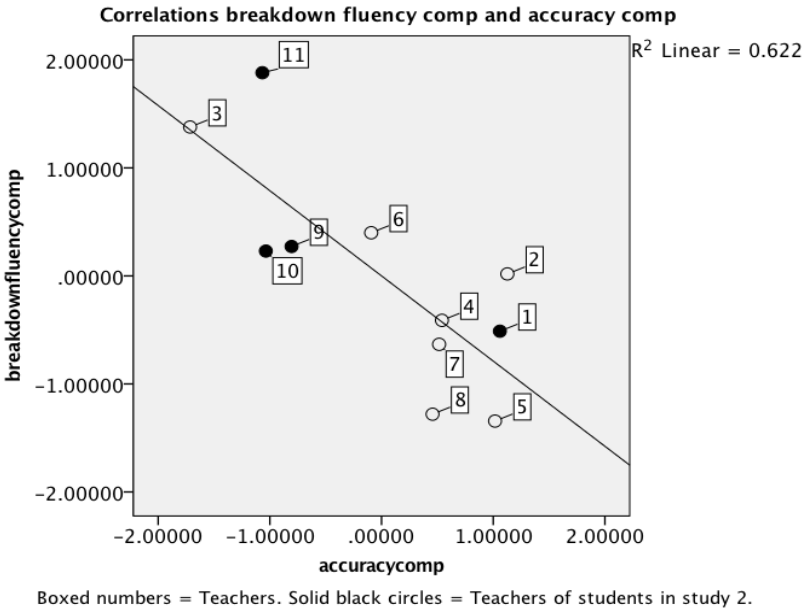

Figure 38 Scatterplot of teachers' breakdown fluency and accuracy

There was a significant negative linear regression between breakdown fluency and accuracy as well ($F(1, 9) = 14.801$, $p = .004$) with an R^2 of 0.622 (*Figure 38*). 62% of the variability in breakdown fluency was explained by the variability in accuracy. When accuracy increased, breakdown fluency decreased and vice versa.

Lexical diversity was significantly correlated with all three fluency subdimensions (Table 23), but not with accuracy and complexity (Table 22).

Table 23 *Correlations lexical diversity and fluency*

		Speed fluency comp	Breakdown fluency comp	Repair fluency comp
Lexical diversity D vocd	r_s	.673*	-.700*	-.664*
	p	.023	.016	.026

$N = 11$
*p < .05. (2-tailed).

Lexical diversity and speed fluency were significantly positively correlated, indicating that the more lexically diverse, the faster the produced language and vice versa. Lexical diversity and breakdown fluency as well as repair fluency were significantly negatively correlated. The more lexically diverse, the fewer breakdowns and the fewer repairs were recorded in the language production. Simple linear regression analyses showed lexical diversity as a significant predictor of speed fluency (p = .03) and breakdown fluency (p = .027), but not significant for repair fluency (p = .065) (Appendix FF).

To sum up, the correlation analyses showed significant correlations between complexity and accuracy, complexity, breakdown and speed fluency, as well as between accuracy, breakdown and speed fluency. Regression analyses indicated that complexity, accuracy, and fluency in terms of speed and less breakdown significantly predicted one another. Lexical diversity showed significant correlations with all subdimensions of fluency, but not with complexity and accuracy.

3.4.1.2 Discussion of Study 3 Part I: Principal Components and CAF Relations

Research question RQ5 investigated how the CAF dimensions can be transformed into a scale that can be used for further analyses. Research question RQ6 asked how complexity, accuracy, and fluency in the teacher's L2 performance relate to one another. It was predicted that all three CAF dimensions correlate (H6).

The analysis of the CAF measures showed a clear picture of the interrelationships and the applied measures of all three dimensions of linguistic performance. In the following, the individual measures and their relationships are discussed first. Second, the relationships between the CAF dimensions are evaluated.

Each dimension was first analyzed according to their underlying measures. This was a necessary step to examine the relationships between the measures and their contributions to each CAF dimension. The calculations of the measures underpinning each CAF dimension lead to the following results: All the syntactic measures for complexity correlated and contributed to the complexity dimension. Lexical diversity as measured in vocd did not correlate with the syntactic measures. Therefore, the current results suggest supporting an approach that considers lexical diversity as measured in vocd its own dimension of language performance that may not be conglomerated with syntactic complexity, as it has commonly been in CAF research (see 2.1.4.2). To this date, extending the CAF framework into a CALF framework has not yet taken place consistently in research, but studies are beginning to base their analyses on an extended CALF framework (e.g. Tavakoli, 2018).

The measures underpinning accuracy significantly correlated and contributed highly to the accuracy dimension. The fluency measures were divided into the subdimensions speed, breakdown and repair fluency. The measures underlying speed fluency all showed significant correlations with at least one other speed fluency measure. In addition, they highly contributed to the subdimension speed fluency, as was shown in the PCA. Of the measures calculated for breakdown fluency, only filled pauses per minute did not show any significant correlation with any of the other breakdown fluency measures. Similarly, the PCA results showed that filled pauses did not contribute significantly to the subdimension breakdown fluency. The two measures for repair fluency – repairs per minute and repetitions per minutes – both contributed highly to the subdimension repair fluency. However, repair fluency did not load significantly on the overall performance.

It was possible to condense the measures underlying the CAF dimensions in a Principal Component Analysis, which was the research question asked in RQ5. The correlations and regression analyses between the composite scores for each dimension revealed several significant relationships, about which RQ6 asked. Lexical diversity correlated significantly with speed fluency, breakdown fluency, and repair fluency. There was no significant correlation with the ratio of filled pauses. Also, no significant correlation was found between lexical diversity and accuracy or complexity.

The findings considering the correlations between lexical diversity and the three fluency coefficients of speed, breakdown, and repair fluency indicated the following: The more diverse the vocabulary, the predictably faster the speech and the fewer breakdowns. Even though correlations do not necessarily signify causality but only a relationship, it is reasonable to suggest that a speaker who has acquired a great range of words will face fewer instances of lexically induced pausing, faster speech, and fewer repairs. Indeed, the regression analyses showed that lexical diversity significantly predicted higher speed and lower breakdown fluency, but not repair fluency. Lexis can therefore be assumed to relate with fluency in terms of breakdown and speed: Lexical retrieval might influence fluency in such a way that it can promote or hamper fluency. This finding supports previous studies that found a positive relationship between fluency and lexical diversity (e.g. Vercellotti, 2015).

In return, the findings also suggest that retrieving lexically diverse items, which may be expected to be more challenging to process, did not appear to result in slower speech, more and longer pauses, or more instances of repair. A similar point has been made by Vercellotti (2015) as well, whose longitudinal data showed that lexical diversity was correlated with fluency. The present findings suggest that automaticity in language production may

be assisted by a diverse productive lexicon and result in more fluency in terms of faster speech with fewer instances of breakdown.

Filled pauses, however, were not correlated significantly with lexical diversity in the present study. A less diverse vocabulary neither related with more filled pauses nor were fewer filled pauses correlated with a higher lexical diversity. This finding indicates some support for research that found the use of fillers a feature of individual speaking style that can also be found in a speaker's L1, as mentioned in section 2.1.5 (e.g. N. H. de Jong et al., 2015). Filled pauses therefore do not seem to be informative of any processing and speech automaticity that are specific to the L2 production of the speakers in the present study. This may also imply that rating a speaker's L2 proficiency cannot validly include the *uhms* and *errs* as indicators of dysfluent L2 speech, unless the speaker's L1 pausing behavior is included as well. Future research may aim at more comparative studies of L1 and L2 linguistic performance of individual speakers.

The present findings add to the current state of the art of research in the CAF framework, in which studies have examined correlations between the CAF dimensions as well as the effects of a particular factor, for example task design, on each of the dimensions (chapter 2.1). Concerning the relationships between the CAF dimensions as calculated in the composite scores, the results showed significant mutual relationships between complexity, accuracy, and fluency. These findings indicate that none of the dimensions came at the expense of another: When the speakers scored comparably high on accuracy, for example, they also scored high on fluency as well as complexity respectively. The dimensions were mutually interrelated. Regression analyses supported the interrelationships by revealing significant regressions between the dimensions. Hypothesis H6 was supported, as correlations were found between the composite scores of complexity, accuracy, and fluency in terms of its subdimensions of breakdown fluency and speed fluency.

In light of previous studies as discussed in section 2.1.6, the results may add to the discussion of Skehan's (1998) Limited Capacity Hypothesis and trade-off hypothesis (Skehan, 2009), which predict L2 speakers to be restricted in their control over complexity, accuracy, and fluency in a way that they cannot cater to all of the dimensions simultaneously. As discussed, trade-off effects need to be considered within the frame of the particular research design. For the participants in the current study, a trade-off between the CAF dimensions was not found. Instead, the present findings suggest that the group of L2 speakers in this study drew on all the dimensions simultaneously while speaking and did not show to attend to any one particular performance dimension in performing this particular cross-sectional study's task.

In addition, Robinson's Cognition Hypothesis (2011) assumes tasks to promote either fluency, or complexity *and* accuracy, which also relates to Skehan's (1998, 2009) idea of meaning-over-form in second language production – meaning relating to fluency, and form referring to complexity and accuracy (see section 2.1). According to the Cognition Hypothesis, two of the dimensions, namely complexity and accuracy, are therefore more closely connected in the participants' speech production and increase jointly with task complexity.

The present findings could not test the effect of increasing task difficulty, as the interview study was cross-sectional. Thus the study did not involve change over time and therefore does not make any claims about the development of the CAF dimensions. Nor was the interview study set out to examine task effects, but instead aimed to limit possible task effects by keeping consistency in the task, the task format, and the task conditions for the participants. The study can therefore not make any suggestions about how different tasks or task conditions affect the CAF performance.

Nonetheless, the current results showed a stronger relationship between the two dimensions accuracy and complexity than between any other two dimensions: 81% of the variance between accuracy and complexity was explained by the correlation between those two variables, whereas the variances between the remaining combinations between breakdown and speed fluency, and complexity and accuracy respectively were all between 62% and 67%. In the present results, there was a stronger linkage between accuracy and complexity, but no trade-off effects were observed between the dimensions in performing the task in the study. Instead, the dimensions covaried.

Robinson's (2011) hypothesis suggests that the stronger linkage between complexity and accuracy comes into effect with increasing task difficulty. The current studies did not examine the speakers' performances on different tasks since the current study was not concerned with change in speakers' performances and aimed at limiting task effects. While conducting the interviews, however, the participants showed different degrees of easiness with keeping the conversation in the interview flowing. Thus, the interviews may have been more demanding for some participants than for others, who showed few difficulties in understanding and actively participating in the interview. So even though the task format of the interview was alike for all participants, it may have been different in terms of how challenging the participants perceived the task. In fact, as was pointed out by Vercellotti (2015), cross-sectional studies on CAF have rarely looked at different proficiency levels when investigating trade-off effects. She also notes that different proficiency groups can be used to represent development in cross-sectional designs. Future research could incorporate how challenging a task

is perceived by the participants in order to relate language performance and subjective difficulty, if a cross-sectional study design is chosen.

Since the participants in the current study varied in their performance and were not of a homogeneous performance group but scattered along the regression line of each of the bivariate CAF relations illustrated in the scatterplot graphs (*Figure 34* to *Figure 38*), the present results also give no indication that different performance levels of the L2 force speakers to focus on different performance dimensions. The CAF relationships did not indicate being dependent on the speaker's individual state of L2 language development. The scores for Teacher 3, for example, were comparably low on all the CAF dimensions, while those teachers who scored relatively high (e.g. Teacher 5) also scored comparably high on all the CAF dimensions.

There were significant correlations between each of the CAF dimensions, regardless of how high or low the speaker ranged among the group of participants, as was illustrated in *Figure 34* through *Figure 38*. Thus, there is no indication in the present results that the capacity to draw on each dimension of language performance during the process of language production is notably different between speakers at comparably lower levels, for whom the task might have felt more difficult, and higher-level speakers.

Since the interview data was based on a cross-sectional research design and the teachers' L2 development over time was not reported, neither supporting nor contradicting evidence can be stated of whether a different task would elicit different CAF performance, which has been suggested by some studies on the relationship of tasks and CAF performances (e.g. N. de Jong & Vercellotti, 2015; Foster & Tavakoli, 2009; Michel et al., 2007), as shown in section 2.1.5. As the task in the present study was the same for all participants in order to establish similar data elicitation procedures, the effect the task itself could have had on the performance and on results differing between the participants can be expected to have been minimal. The present findings seem to answer Vercellotti's question: "Certainly, speakers may focus on one component, but must they?" (2015, p. 1). The participants in the current study did not have to.

As Vercellotti (2015) comments with respect to the trade-off hypothesis, "[a]ctual trade-off effects should be detected at the individual level because trade-off effects are hypothesized to be exerted within the individual" (p. 17). The interrelationships found in the current study between complexity, accuracy, and fluency of the participants can be summed up as showing no trade-off across the group.

According to the findings resulting from the present study, the implications Robinson's (2011) Cognition Hypothesis as well as Skehan's (1998) Limited Capacity Hypothesis, or Trade-Off Hypothesis (2009), have for second language teaching, may not be as straightforward as expected.

Skehan (1998) argued that teaching has to take the trade-off effects into account and drive learner's attention "to focus on particular aspects of language performance" (p. 288). Such focusing can be found in teaching approaches that suggest activities gearing at complexity, accuracy, or fluency individually, as presented in Thornbury (2000), for example. Considering the CAF performance of the participants in the present study, however, such a focus may not always be necessary since the CAF dimensions showed to co-vary. Longitudinal data on the development of the CAF dimensions over time have shown similar support for learners' ablility to focus on all three dimensions without task-based curricula being manipulated towards a single dimension. Vercellotti (2015, p. 18) therefore suggests her findings "challenge th[e] recommendation" made by Skehan (1998) to manipulate learners' attention to a particular CAF dimension.

The present results indicate that if complexity, accuracy, and fluency can be performed simultaneously, teaching approaches may include activities that demand language production in all the dimensions. Likewise, target language training of future L2 teachers may aim at advancing language performance on all three dimensions simultaneously instead of focusing on one at a time.

3.4.2 Study 3 Part II: Teacher Performance and Students' Results

As was discussed in section 2.2.5, the language performance of teachers or other target language providers – as part of an overall language proficiency – may have a beneficial effect on children's language acquisition in many ways. To briefly repeat: First, features in the linguistic input of caretakers were found to correlate with children's language acquisition of those very features (section 2.2.4). Second, over-all target language proficiency was expected to affect teachers' confidence, spontaneity in using the language, and delivery of linguistically diverse input (section 2.2.5.2). Third, over-all target language proficiency was assumed to foster teaching strategies that may assist children in acquiring a second language (section 2.2.5.3).

The following sections analyze the relations between four of the teachers' CAF performance and the students' grammar and vocabulary test results. The principal component scores obtained in the previous synthesis study Part I for each CAF dimension were the base for the regression analyses carried out in the following Part II between the teachers' CAF performance and the students' test scores. Part II connects the PCA results, the four teachers' CAF performance, and their respective students' test results.

The research questions and according hypotheses are as follows: (RQ7) How does the teachers' L2 English performance, as measured in complexity, accuracy, and fluency, relate to their students' L2 receptive vocabulary and grammar development? It was hypothesized that there was a positive relationship between the teachers' CAF performance as well as each of the CAF dimensions and the students' receptive grammar and vocabulary development (H7).

(RQ8) If there is a relationship between teachers' L2 performance and children's foreign language acquisition, is there an additional effect by the classroom L2 use as rated by the teachers? It was hypothesized that the teachers' adapted L2 use in the classroom would moderate a possible CAF effect on the children's receptive grammar and vocabulary development (H8).

3.4.2.1 Data Analysis of Teacher Performance and Students' Tests

For each CAF dimension, the principal components calculated in the PCA were condensed to a composite factor score to arrive at a computable overall measure for each dimension (see section 3.4.1). A Spearman's rho test of correlations was performed to relate the principal components coefficients and the students' vocabulary and grammar test results at time one as well as time two. In addition, the relation between the principal components coefficients and the difference between the vocabulary and grammar test scores at time 1 and time 2 was computed to analyze whether the teachers' CAF performance was related to the development of the children's receptive skills between the two test times.

As computed in Part I above, the principal components scores were *speedfluency comp, breakdown fluency comp, repair fluency comp, accuracy comp,* and *complexity comp.* In addition, *lexical diversity D vocd* was included as its own variable, since it was not significantly correlated to any complexity measures, as shown in the Spearman's rho correlation results in section 3.4.1.1.1.

The following section reports the results of the relationships between the teachers' CAF scores and the students' test scores. Spearman's rho test of correlation was computed to detect any significant correlations between the four teachers' CAF performance and the students' development in the vocabulary and grammar scores between time one and time two. Finally, multiple regression analyses were performed to compute whether and which CAF dimension predicted a gain in the students' receptive development.

Teacher classroom language behavior was not observed within the scope of the current studies. However, teacher-talk may be modified in the classroom, as was discussed in section 2.2.5.1. Therefore, the answers of the teacher questionnaires of the substudy in 3.2.5 on the amount of language modification in the classroom were integrated in a final calculation as

additional variables to control for, after completing all calculations on the CAF dimensions, their internal relationships, and their correlations with the students' test results.

3.4.2.2 Results of Teacher Performance and Students' Tests

Table 24 presents the results of the correlation analyses between students' vocabulary as well as grammar development, as represented in the difference between the scores at time one and time two, and the CAF composite scores.

Table 24 *Correlation teachers' CAF and students' test results (t2-t1)*

	Vocabulary difference		Grammar difference	
	r_s	p	r_s	p
Speedfluency comp	-.020	.892	-.392**	.000
Breakdown fluency comp	.020	.892	.392**	.000
Repair fluency comp	.001	.993	.149	.169
Accuracy comp	-.017	.906	-.312**	.003
Complexity comp	-.020	.892	-.392**	.000
Lexical diversity D vocd	-.020	.892	-.285**	.008

Note. Identical correlation coefficients result from ranking the teacher subsets of three (vocabulary) and four (grammar) teachers. The teachers' individual raw scores for each measure varied.

Vocabulary test n = 49, Grammar test n = 87.

*p < .05. **p < .01. (2-tailed).

The grammar development correlated significantly with each CAF component score, except repair fluency. Vocabulary development did not show any significant correlation with any of the CAF dimensions (Table 24). The students' development in receptive grammar revealed significant negative correlations with the teacher's speed fluency, accuracy, complexity, and lexical diversity. These were unexpected findings, as a positive development in student grammar scores correlated with lower teacher accuracy, lower speed fluency, lower complexity, and lower lexical diversity. There was a significant positive correlation between the students' grammar difference and the teachers' breakdown fluency ($r_s(85) = .392^{**}$, $p < .001$). Put differently, the students' improvement in grammar correlated with the

teachers' higher number and longer lengths of pauses. Repair fluency was not significantly correlated with the grammar difference.

Additional Spearman's rho correlation analyses were computed to distinguish between the vocabulary and grammar scores at each of the test times separately. Table 25 shows the correlations between the composite scores and vocabulary and grammar at time 1 and time 2. There were no significant correlations between the CAF composite scores and the vocabulary scores at either time. There were significant correlations between all the CAF composite scores except lexical diversity and the scores for grammar.

Table 25 *Correlations teachers' CAF and students' tests at t1 and t2*

		Vocabtest t1[a]	Vocabtest t2[b]	Gramtest t1[c]	Gramtest t2[d]
Speed fluency comp	r_s	-.085	-.021	.068	-.371**
	p	.553	.881	.514	.000
Breakdown fluency comp	r_s	.085	.021	-.068	.371**
	p	.553	.881	.514	.000
Repair fluency comp	r_s	.172	.129	.049	.216*
	p	.227	.346	.640	.028
Accuracy comp	r_s	.095	.124	.047	-.249*
	p	.508	.366	.654	.011
Complexity comp	r_s	-.085	-.021	.068	-.371**
	p	.553	.881	.514	.000
Lexical diversity D vocd	r_s	-.085	-.021	.191	-.183
	p	.553	.881	.065	.062

Note. Identical correlation coefficients result from ranking the teacher subsets of three (vocabulary) and four (grammar) teachers. The teachers' individual raw scores for each measure varied. [a]$n = 51$. [b]$n = 55$. [c]$n = 94$. [d]$n = 104$. Numbers indicate all completed tests.

*p < .05. **p < .01. (2-tailed).

How much of the variance in the grammar difference can be explained by the CAF scores and possibly the teacher ratings, was examined next. For two reasons the difference scores were used: Only students who took the tests at both times were included in the difference scores. The difference scores were considered to determine the effect of the teachers' linguistic performance on the students' grammar development.

Multiple regression analyses were conducted with grammar difference as the dependent variable and the CAF composites as independent variables to predict the students' difference in the grammar scores.[28] The composite scores of *complexity comp, accuracy comp, speed fluency comp, repair fluency comp, breakdown fluency comp*, and *lexical diversity* were entered into the model as independent variables.

Only breakdown fluency remained a significant predictor of the difference in the grammar scores ($\beta = .348$, $p = .001$) (Table 26). The regression model showed that 12.1% of the variance in the grammar difference was accounted for by the teachers' breakdown fluency.

Table 26 *Multiple regression results for grammar difference*

	Grammar difference t2-t1				
Variable	*B*	*SE B*	*β*	*t*	*F*
Breakdown fluency comp	3.141	.918	.348**	3.423	11.716**

Note. $R^2 = .121$. (** $p < .001$)

A significant regression coefficient was found for breakdown fluency ($F(1, 85) = 11.716$, $p = .001$) with an R^2 of .121 (adjusted $R^2 = .111$), indicative of a moderate goodness-of-fit.[29] Participants' predicted grammar difference was equal to 1.109 + 3.141, with breakdown fluency comp measured in pausing. The grammar difference increased by 3.1% for each instance of breakdown fluency. Breakdown fluency remained the only significant predictor of grammar difference in the model. The regression model excluded the composite variables speedfluency comp, repair fluency comp, accuracy comp, complexity comp, and the over-all performance variable as non-significant predictors (Appendix GG).

As breakdown fluency is a dimension based on pauses, a positive breakdown fluency value indicates less fluent language production in terms of the pausing. Thus, breakdown fluency as a significant predictor of change in the receptive grammar scores indicates more extensive pausing in a model speaker's language to have a positive effect on receptive grammar development.

An additional calculation was performed to analyze if the teachers' rated in-class target language use, as rated by the teachers in the questionnaires

28 Stepwise regression allows the predictors to be entered simultaneously in a non-hierarchical fashion, based on mathematical criteria only (Field, 2013).

29 According to Cohen's benchmarks .02 (small), .13 (medium), .26 (large). Quoted in Field (Field, 2013).

(section 3.2.5.2), moderated the effect found for breakdown fluency. A regression analysis was run with grammar difference as the dependent variable, breakdown fluency as the independent variable, and the controlling variables adaptive language score, score of perceived language proficiency for teaching, and the score of perceived speaking proficiency. Table 27 shows that there was no significant additional effect of the teachers' self-rated adapted language in the classroom on the development of grammar.

Table 27 *Regression results breakdown fluency, teachers' rating as controlling variables*

	Grammar difference t2-t1				
Variable	*B*	*SE B*	*β*	*t*	*F*
Model 1 Breakdown fluency comp	2.947	.971	.335*	3.034	9.208*
Model 2 Breakdown fluency comp Adaptive scale	3.162 1.476	1.028 2.241	.359* .077	3.076 .659	4.785*

Note. R^2 Model 1 = .112. R^2 Model 2 = .117 (* $p < .005$)

Neither the adaptive language score, the score of perceived language proficiency for teaching, nor the score of perceived speaking proficiency significantly moderated the effect of breakdown fluency on grammar differences ($p = .512$). The effect of breakdown fluency remained significant even when the teacher questionnaire variables were controlled for ($p = .003$) and still explained 11% percent of the variance in the development of receptive grammar (Appendix HH).

In sum, these analyses converged on indicating that students' receptive grammar development was related to the teachers' breakdown fluency, which is based on the teachers' use of pauses while speaking.

3.4.2.3 Discussion of Study 3 Part II: Teacher Performance and Students' Tests

To refer back to the original research questions and hypotheses, they are repeated as follows: Research question RQ7 asked how the teachers' English performance, as measured in complexity, accuracy, and fluency, related to their students' receptive vocabulary and grammar development. It was predicted that there was a positive relationship between the teachers' CAF performance as well as each of the CAF dimensions and the students' receptive grammar and vocabulary development (H7).

Research question RQ8 asked whether there was an additional effect of the classroom L2 use as rated by the teachers in case there was a relationship between teachers' L2 performance and children's foreign language acquisition. It was hypothesized that the teachers' adapted L2 use in the classroom would moderate a possible CAF effect on the children's receptive grammar and vocabulary development (H8).

The results of the relationships between the teachers' performance and the students' development of their receptive vocabulary and grammar suggest a complex relationship between the linguistic performance of the model teacher speakers and the second language acquisition of their students, the elementary school children in the present study.

The study synthesis of the teachers' performance and their students' receptive vocabulary and grammar development focused on the linguistic performance of the teachers as a singled-out factor in second language learning. As such, no direct positive relationship between teachers' higher linguistic performance and the students' progress in vocabulary and grammar was reported.

Thus, a comparably high performance on the CAF dimensions did not come with comparably high scores in the vocabulary and grammar tests of the students, neither in terms of the attainment at each test time nor the rate of development between both times, as Teachers 1 and 11 showed. As was illustrated in the scatterplots showing the teachers' CAF performances (*Figure 34* to *Figure 38*), Teacher 1 had comparably higher CAF composite scores than the other three teachers, but teacher 11's students showed a significant gain in receptive grammar. Teacher 1's students scored lower on the grammar tests in terms of the development between times one and two as well as the grammar attainment at time two. Teacher 11, on the other hand, scored comparably low on the CAF composite scores, but her students scored significantly higher on the grammar test than Teacher 1's students. Hypothesis H7, which predicted higher teacher CAF performance as well as each of its individual CAF dimensions to relate positively to their students receptive L2 development, was therefore rejected.

In his compilation of findings on non-native speaker teachers and proficiency, Llurda (2006) states that a high level of proficiency was a critical condition for language teachers but one "which alone does not guarantee successful language teaching" (p. 234). Additional participant information on the teachers revealed that the teachers' background in terms of holding a degree in English and having staying abroad did not affect the results: Unlike Teacher 1, Teacher 11 had no degree in English and had not lived in an English-speaking country. Teacher 11 was, however, more experienced than Teacher 1 in terms of the years this teacher had been teaching. In addition, Teacher 11 held a degree in a different language, which may have had a beneficial effect.

The picture of the effect of teachers' L2 performance on children's language acquisition may change if internal variables, such as cognitive factors, the children's motivation, and their L1, or if different external variables such as methodological approaches were included as covariates of the teachers' linguistic performance. Students may be able to benefit more from the teachers' linguistic performance if it is combined with additional variables that promise to have a beneficial effect on language acquisition in general. However, it has been shown that it is an underexamined area of research how teachers' proficiency relates to assisting strategies such as the factors summarized by Kersten (2019) (section 2.2.5.3) and similarly by VanPatten and Benati (2015) for modified speech (section 2.2.5.1). The vast majority of studies supporting beneficial teaching strategies were based on native-speaker teachers or teachers in bilingual and immersion contexts, who may be considered more proficient in the L2 than untrained, regular school English teachers. As has been pointed out (section 2.1), the much more common situation for children acquiring English in instructed learning settings such as at regular elementary schools in Germany is being exposed to a limited amount of English by teachers who are most often not trained in English. Corrective feedback, as an example of assisting teaching strategies, may include erroneous feedback, whose effect cannot be studied if it is not operationalized in terms of its linguistic properties as well. As has also been pointed out, it remains subject to further investigation how promoting diverse and complex input goes along with modified input without contradicting one another.

It could be the subject of future research that aims to examine the interrelationships of diverse variables at interplay, to take into account the teachers' linguistic performance as part of the target language proficiency as well. Certainly, as Chambless (2012, p. s154) remarks, an additional approach to examining how teacher proficiency, teaching effectiveness as observed in supportive teaching strategies, and students' L2 learning relate, is only possible with a great number of coordinated research teams applying a multitude of qualitative and quantitative methods. This could improve the ecological validity in terms of the relationships between teachers' linguistic L2 performance as a determiner of an overall L2 proficiency, their classroom behavior, and their students' L2 development.

Thus, the findings in the synthesis call for future research that takes the teachers' linguistic performance into account alongside a variety of other possible influencing factors. Favorably, those factors would include internal individual factors of the students, but also external factors. In terms of the schools studied here, it was notably the students of the largest class and school whose means improved the least, even though their teacher (Teacher 1) performed comparably high on the CAF dimensions. Teaching circumstances such as the size of the school, the student body as well as the

class size may therefore be additional influential factors to take into account in future research on which factors may have an impact on children's second language acquisition in a learning environment with little exposure to the target language.

Another factor might also have had some impact on the results: The students are still at an initial stage of second language learning. One explanation for not finding a positive correlation between a teacher's high performance and the students' target language development could be that at the emergent stage of second language acquisition, at which the students in this study were, the teacher's linguistic performance alone is in fact a comparably weak factor at this amount of input, when the teacher's linguistic performance is within the medium range of low and high L2 English-speaking teachers. Within the entire group of the eleven teachers interviewed and analyzed for their CAF performances, the subset of the four teachers indeed ranged in the middle, as was illustrated in the regression graphs in section 3.4.1.1.5. This would also be in accordance with Unsworth et al.'s (2015) findings, which indicated that the children taught by a CEFR level C1 proficient teacher, or at least a native speaking teacher teaching alongside a lower proficient L2 teacher, achieved higher tests scores. This may suggest that a linguistic performance threshold needs to be met in order to have a beneficial effect on the acquisition of English at elementary level schooling. The findings of the present study showed that in fact the higher CAF scores correlated negatively with the students' scores in grammar, yet only one subdimension emerged as a significant predictor of the grammar improvement. Additional data on students taught by teachers who perform lower on the CAF dimensions could provide more insights as to how the L2 of those students develops.

The effects of a higher teacher L2 performance may also play out positively at an advanced stage in the development, which would not show until a much later time. Occasionally, such a late long-term effect has been observed: Teachers' language proficiency predicted grammar scores only at the second posttest in Unsworth et al.'s study (2015). Their finding suggests that teacher proficiency effects can take more time to show. Similarly, Aukrust (2007) found that diverse and complex native-speaker input predicted the preschoolers' outcomes in their second language, which was Norwegian, two years later. Jaekel et al. (2017) also found long-term effects years later.

The current study covered the last year of elementary school and could therefore not calculate a possible effect at a later stage than investigated. Further research could also show if the relationship between teacher performance and student outcomes differs at a later stage of second language acquisition, when the students are possibly more susceptible to features in the linguistic input and are able to benefit from more complex, more

lexically diverse, and more fluent language of the model speakers. At the same time, such research could illuminate Ellis and Shintani's (2014, p. 189) stance that only lower proficiency learners may benefit from simplified input, while it is not considered adequate input for learners who are more advanced.

Language contact intensity in terms of the amount of exposure was another aspect that predicted a gain in receptive L2 vocabulary or grammar scores in a number of studies (e.g. Kersten, Schüle, et al., forthc. Lightbown, 2014; Maier et al., 2016; Muñoz, 2014; Rohde, 2010; Saito & Hanzawa, 2018; Steinlen et al., 2010; Unsworth, 2016a; Weitz et al., 2010). As one of the few studies integrating L2 proficiency in their study, as discussed in section 2.2.5.2, Unsworth et al.'s (2015) study has provided insights into the interrelation of proficiency and contact time with the second language: Contact time as well as teacher proficiency positively related to an improvement in vocabulary and grammar scores, but only if the teacher's proficiency level was higher than CEFR B1 or if a native speaker teacher taught alongside, and only when the amount of instruction time was higher than 60 minutes per week – either 60 to 120 minutes or more than 120 minutes. The hours of instruction of the present studies were the same mandatory load of two 45-minute lessons per week for all the groups in the study. Input intensity, as defined in the amount of instruction time, could therefore not explain any differences between the groups in the current study. However, the amount of weekly English contact time may have been too low, which could explain why the higher linguistic performance of the teachers in the present study was not able to predict an advance in vocabulary and grammar.

It is possible that the amount of instruction of two lessons per week was not sufficient to show any effect of teachers' higher language performance. For example, the age effect Larson-Hall (2008) studied, played out only when the input amount was over four hours. Similarly, teacher language performance may come into effect differently at a higher amount of instruction time.

As it was not within the scope of the present study, it remains subject to future research whether particular teaching strategies as discussed in section 2.2.5.3 had any effect on the development of the students' receptive grammar and vocabulary development. For example, examining the accuracy dimension in the teachers' performance and its relationship to the in-class corrective feedback of lower-level teachers would elucidate if errors in the input are perpetuated in corrective feedback or not and how such a relationship could affect the children's second language acquisition.

A cautionary note needs to be reiterated regarding the comparability of the current study and other studies. Keeping in mind that the children for example in Aukrust's (2007) study as well as those in Unsworth et al.'s (2015) study were younger than the students of the present study, comparability is

limited. In addition, the learning environments were different, as Aukrust's subjects were second language learners in a second language environment, and Unsworth et al.'s subjects were likely to have been more exposed to the target language outside of the classroom, to name only a few outstanding differences. Additional exposure to English was not expected for the children in the current studies (see section 3.3.1.1), even though contact with English through possible internet usage could not be ruled out.

One of the merits of conducting research in a school setting of a country with very limited presence of English in other fields of everyday life, such as TV broadcasting, was its controlled setting of foreign language classroom instruction, in which the teacher represented the students' primary and often only contact to the language. A few studies have shown that the presence of outside-of-class English has a positive effect in children's second language acquisition (e.g. Enever, 2011, 2014; Huang et al., 2018; Kuppens, 2010; Lindgren & Muñoz, 2013), as discussed in section 2.2.6. More studies in similar contexts to the current one that include participants of the same age range and who are exposed to an equivalent intensity of the L2, could enhance comparability among studies and eventually shed more light on the specifics of second language acquisition in a non-naturalistic environment.

Another reason why there was no gain in some student groups, even when the teacher scored comparably high on the CAF dimensions, may have been that the linguistic input of the respective teacher was too advanced for the students to process and to take in – in the sense of *intake* in Gass's (2018) model (see 2.2.2). This explanation would be in accordance with the idea of comprehensible and comprehended input being a prerequisite of second language learning. It would also be in line with those theoretical models and hypotheses presented in chapter 2.2 which postulate comprehensible (e.g. Krashen, 1981, 1985, 2009) or comprehended input (e.g. Gass, 2013, 2018; Gass & Selinker, 2008) as a crucial issue in the requirements of second language acquisition. Whereas higher complexity and fluency may be reasonably explained to correlate with lower test scores, because the target language may have been beyond the students' capacity to process the linguistic input to such an extent that higher complexity and fluency did not bring about any benefits, accuracy then appears in a different light. At first glance it may seem contradictory that higher accuracy in the teachers' language was negatively correlated with the students' development of receptive grammar. Since all three CAF dimensions correlate significantly with one another, however, accuracy cannot be viewed independently: it co-varied with complexity and fluency and will likewise increase as well as decrease with complexity and fluency.

What is striking about the relationship between accuracy and the development of receptive grammar is the fact that according to these findings, lower accuracy did not *hamper* the development of grammar at

this stage. This would suggest that children learning a foreign language can handle erroneous linguistic input – an idea underlying theories of Universal Grammar, which explain children's ability to deal with faulty linguistic input by their innate capacity to acquire any language, even though the input children are exposed to is not error-free at all times.

Yet it seems ambiguous that lower accuracy in the teachers' language would correlate significantly with greater increase in the receptive grammar of the students. Apart from the idea that children seem to be able to compensate for possible erroneous linguistic input, which is the case in first language acquisition as well, another factor might be at play in the cases studied here: Since the teachers share the same first language with a great majority of the children, L1 transfer can be expected to be not only similar, but also not impeding comprehension – or in fact facilitating understanding. Errors resulting from L1 transfer will not only be similar but also be comprehended more easily when the interlocutors share the same first language. A similar idea was presented by Loder Buechel (2015), who did not find the teachers' proficiency in her study positively related to their elementary school students' L2 English attainment either.

Plausibly, errors might then not be recognized by either party, neither the teacher, nor the students. An utterance like *What is this for a door?*[30] asked by an L2 English-speaking daycare teacher pointing at a closed door, can cause some ambiguity in native English speakers and non-German speakers: the question could mean *What is this door for?*, which would be grammatically correct but highly unlikely to be asked if the speaker wants to know what is behind the door. However, if the interlocutor literally translates the utterance back into German, the daycare teacher's first language, the utterance becomes *Was ist das für eine Tür?* The erroneous English question transforms into a correct German construction with a clear aim to receive information on where the door opens to. Thus, an erroneous construction like *What is this for a door?* could in fact help the German L1 children understand the question. Non-German native English speakers might have been confused in this case.

The results for the development of vocabulary did not show a statistically significant correlation with the teachers' lexical diversity. Yet it is remarkable that those students whose teacher showed the highest lexical diversity (Teacher 10), showed the most increase in the mean receptive vocabulary raw score between the two test times, although lexical diversity did not show as a predictor of vocabulary development. The mean gain in the vocabulary scores was not significant. However, there may still be a relationship between lexical diversity and vocabulary development, even

30 Personal encounter.

though such a relationship was not detected as statistically significant in the current results.

Two more observations do not seem to support the explanation that the teachers' performance might have been too advanced for the students to process and make use of: After all, the four teachers in the current study reported simplifying their language in the classroom by pausing, using a less complex sentence structure, repeating words and phrases, and using simpler vocabulary. In addition, among the eleven teachers interviewed, the subset of the four teachers whose classes were tested were not among the highest-scoring teachers, although Teacher 1 often ranged at the higher middle ends of CAF performance.

The teachers' questionnaires yielded some additional insights to answering research question RQ8, asking whether there was an additional effect of the amount of language modification in the classroom. The teachers did not differ significantly in their estimation of how much target language they spoke in class. Thus, there is no indication that the differences in the students' development of receptive scores can trace back to a significant variance in the amount of English spoken by the teachers in class. In addition, the questionnaire results did not indicate that modification or simplification in the teachers' language had any effect, as was shown in the additional regression models into which the variable *adaptive language* was included as a moderating variable. Therefore, hypothesis H8, which predicted an additional effect of adapted language in the classroom, was refuted. As has been pointed out in the teacher questionnaire substudy (section 3.2.5), however, questionnaire data based on self-reports remain limited in their validity and reliability. Yet the most reliable items may have been the self-reports on the amount of language modification in the classroom, as the answers do not necessarily presuppose an expected best answer in such a way that more pausing in the input, for example, is to be aspired or avoided.

Another explanation for not finding any statistically positive effect of higher CAF performance scores on the development of the children's receptive grammar and vocabulary skills may be that the group of teachers investigated here were either too similar in their linguistic performance, or did not go beyond a certain level of language performance. The latter would be in line with hypotheses and models of second language acquisition that suggest not only comprehensible, but also challenging input to be a requirement for developing a second language. As discussed in relation to teacher-talk (section 2.2.5.1), Ellis and Shintani (2014) warned against denying learners diverse linguistic features through simplified input. However, a teacher language performance or proficiency threshold has not been found, even though Unsworth et al. (2015) found higher CEFR-levels of teachers to predict students' L2 outcomes. Studies on teachers

scoring either much lower or much higher than the ones in the present study might illuminate whether there is a threshold of too challenging or too undemanding input and where it could possibly be set. Further research that would look at a larger variety of language performances could clarify if there is no genuine effect of the teachers' linguistic performance or whether there will be one if a particular performance threshold is passed.

Research question RQ7 asked about the relationship between the teachers' CAF performance as well as each CAF dimension and the students' results. The multiple regression analyses were able to shed more light on which particular performance dimension could predict the development of receptive grammar. Of all the CAF dimensions included in the multiple regression models, breakdown fluency evolved as a significant predictor in the gain of grammar between the two times. About 12% of the variance in the grammar difference was accounted for by breakdown fluency. Breakdown fluency remained a significant predictor in the regression models that included all CAF dimensions as well as in those that excluded CAF variables. Therefore, breakdown fluency can be considered the most robust result of the present models and needs to be interpreted with regard to its role in second language acquisition. Breakdown fluency was a significant positive predictor of the grammar development. Since breakdown fluency was based on measures of pausing, a high amount of breakdown indicated low fluency. Hence, hypothesis H7 is refuted, which predicted that there was a positive relationship between the CAF performance as well as each of its dimensions and the students' receptive grammar and vocabulary development. Rather, a specific predictive power of a single dimension was found – breakdown fluency.

As was shown, breakdown fluency is based on the pausing behavior in language production. Pausing can give the listeners time and opportunity to process what is being said, which has been reported to facilitate comprehension. As an isolated dimension in the production dimensions, the relationship between more and longer pauses in the input and the development of receptive grammar could be explained by the positive effect pauses have on the perception of grammar structures.

Time was mentioned as a factor in models of second language acquisition (section 2.2.2). Gass (2018) names time pressure as one of the factors that serve as "input filters" (p. 17), affecting why some input features may be apperceived while others may not. Learners need time to segment the strings of language in the input into manageable units. Gass (2018) suggests that time pressure has a particularly strong effect on processing linguistic input at the early stages of second language acquisition and when the input is predominantly provided orally. In this light, pausing in the input, which was measured in breakdown fluency, gives learners the necessary breaks in the input and thus time to segment words, phrases, and other units. Time

pressure, on the other hand, impedes the learner's opportunity to break the input into manageable pieces.

The positive effect of breakdown fluency on the development of receptive grammar may have been affected by the factor Gass (2018) lists as time pressure in that less time pressure through more frequent and longer pauses in the input benefit the development of grammar. Interestingly, as mentioned in section 2.2.3, Bowers and Vasilyeva (2011) reported a similar result for their study on the growth of receptive vocabulary, when their results indicated that a large total amount of input was positively related to vocabulary growth, but long utterances measured in words per utterance showed a significant negative correlation with vocabulary growth. Concluding, there is some reason to suggest that pausing in the teachers' language performance could give the learners processing time that can benefit the development of receptive target language skills. However, suggestions need to bear in mind possible impacts on the study's ecological validity with respect to the relationship between the teachers' measured linguistic performance and their classroom language use.

The current results as well as Bowers and Vasilyeva's (2011) findings may indicate that with regard to vocabulary development, a large amount of input is necessary. Lexical development was suggested to be highly incidental, which requires a high amount of input (e.g. Laufer & Hulstijn, 2001; Loewen & Sato, 2018; Newton, 2013) (section 2.2.3). This ties to the note of frequency of an item in the input, which according to Gass (2013, p. 500) allows for learners to notice a form and eventually integrate into their language system. Pauses may then be additionally needed to better segment the language input.

Speed fluency, on the other hand, did not significantly predict grammar development in the current study, even though lower speed could be expected to ensure more processing time for the listener as well. However, the results of the regression analyses did not show this relation. This finding may indicate that it is not as much the overall speed of spoken language that buys the learners processing time to deal with linguistic features in the input, but that linguistic data is best processed when it is broken into digestible chunks that are segmented by pauses. Further research may help gain clarification on this particular aspect of breakdown fluency and its relationship with second language acquisition. Breakdown fluency in the input may emerge as a specific feature in the spoken input benefitting its comprehensibility to the learners.

Vocabulary may be processed more independently of the pauses in the input than grammar. As vocabulary acquisition is believed to be more incidental than the intake of syntactic structures, lexical items may enter the storage more easily:

> It is reasonable to assume that the storage component is more likely for vocabulary and smaller chunks of language than for large syntactic strings. This may be due largely to the fact that it is more difficult to hold large bits of language in memory for a long period of time. (Gass, 2018, p. 6)

Therefore, for vocabulary to be processed from the input it may be less dependent on pauses between speech units than grammar. Gass's suggested explanation of what enters the storage section of the acquisition model may explain that no correlation between vocabulary development and the teachers' performance was found in the studies at hand. If words are moved to the storage to be reevaluated later through hypothesis testing in the language acquisition process, it follows that those words will be integrated at least in the receptive vocabulary storage in a more independent way of the teacher's performance and its specific CAF features.

In the current study, however, students' receptive vocabulary did not show a statistically significant increase between the two times of testing. It has been argued in section 2.2.3, referring specifically to DeKeyser (2000), that children may learn to a great extent implicitly at a young age, which requires a large amount of input. Such a sufficiently large amount of input may not have been given for the students in the current study, as they were taught only two 45-minute lessons per week in a non-naturalistic foreign language instruction setting.

Since the development of receptive vocabulary was not statistically significant, an effect of any of the CAF variables predicting vocabulary development could not be computed in the statistical analyses of the present studies. However, the group size may have had an influence on the statistical results, as the group of the students who sat the vocabulary test at both times was smaller than the group who participated in the grammar tests. Larger group sizes are more prone to show statistical significance if there is one to be found.

To sum up, the results do not indicate that the teachers' over-all linguistic performance in terms of their CAF performance positively affected students' outcomes on receptive vocabulary or receptive grammar during their fourth year of elementary school. Instead, breaking down speech by pauses showed to have predictive power in the students' gain of receptive grammar in the current study. As was discussed in chapter 2.2, extensive pausing in the input is part of language providers' input in various context – it is observed in child-directed speech in first language input as well as in teacher-talk in foreign language teaching. Pausing is also considered beneficial for processing language, first as well as second language. In addition, pausing is discussed in research on teaching strategies as an assisting element in teacher input and target language classroom strategies. Therefore, the present result of breakdown fluency as a significant predictor of the development of

receptive grammar is in line with the theoretical assumptions underlying pausing in the input to be beneficial for L2 development.

The current findings demand further studies investigating the relationship between L2 teacher performance and children's second language acquisition. The present findings also suggest examining specific features in the language performance, as they may surface as particularly valuable indicators of performance. Further research examining breakdown fluency as a particular feature among all the features that characterize language input would be illuminating in order to determine its relevance for second language input and its intake.

Several limitations in the study restrict the generalizability of the obtained results: First, the small number of teachers regarding the relation between teacher language performance and student outcomes, second, the large variance among the student scores in all tests at both times of testing, and third, the relatively small differences between the test scores at both times. Replicated studies, studies with more teachers, studies with different tests of language assessment, or studies stretching over a longer observation time span could advance the insights gained from the studies at hand.

4 Conclusions

The present dissertation thesis has investigated how elementary school English teachers performed in their L2 English and how their language performance related to their students' receptive L2 grammar and vocabulary development.

The theoretical background examined several issues of interest to the topic. It was shown that the CAF framework has become a trusted means to measure language performance. However, inconsistencies were detected in the results of studies on language performance, in particular with respect to how findings support hypotheses of whether complexity, accuracy, and fluency can be balanced in second language production, or whether there was an inevitable trade-off between the dimensions.

Linguistic input was then examined in light of how a connection can be drawn between language providers and children's second language acquisition. Research on first language acquisition indicated that children's language acquisition might be affected by their caretakers' linguistic performance. With respect to second language acquisition, a number of hypotheses have been developed that aim to explain which features in the linguistic input assist children in acquiring a second language. The hypotheses are situated mainly in usage-based approaches to second language acquisition, which rest on the assumption that language acquisition is input-driven. Thus, features such as comprehensible input or frequency of forms in the input, but also specifics in teacher-talk are extensively discussed in second language research. However, it was revealed here that research has not yet adapted to the current situation that most English teachers are L2 speakers of English, who may often not be trained in English.

Studies incorporating receptive grammar and vocabulary as indicators of children's L2 language development are growing in number. Standardized tests help to enhance comparability of studies in the field of children's second language acquisition, as results are quantified by standard calculations. However, it has been shown that caution needs to be exercised in comparing and interpreting results, considering the different circumstances in which children learn a second language. A need for more studies on foreign language acquisition in regular elementary schools in a non-target language environment was detected that could cast more light on the nature of early second language development in this particular environment. Input factors can be more clearly isolated in a foreign language classroom setting because confounding out-of-class exposure to the target language can be largely controlled for. This may be even more the case for languages other than English, which are less prominent in media such as the internet.

In order to approach the topic of this thesis empirically, three main studies were conducted. In Study 1, the L2 English teachers' performance in interviews was analyzed according to their complexity, accuracy, and fluency based on a large number of measures. Study 2 examined students' receptive vocabulary and grammar based on their results on the BVPS3 and ELIAS Grammar Test II at two times during the fourth year of elementary school. The synthesis in Study 3 was divided into two parts: Part I analyzed the teachers' CAF results to compute a score for each dimension. A Principal Component Analysis revealed the measures' contribution to each CAF dimension and could calculate component scores. The new composite scores allowed for subsequent analyses of the CAF relationships as well as of the relationship between the teachers' measured linguistic performance and their students' development in receptive grammar and vocabulary.

The teacher interview Study 1 was able to gather a broad range of teachers in terms of their L2 language performance, teaching experience, L2 language training as well as experience with living in an English-speaking country. This was particularly important so as to have variance in the teachers. All eleven teachers revealed mixed results in their English speaking performance. There was variation in all investigated speaking dimensions between the subjects, while the dimension performances within the participants showed similar relations to one another. So speakers who scored comparably high on one CAF dimension also scored high on the other two, and speakers who scored comparably low on one dimension, scored comparably low on the other two as well.

The results of the present thesis have illuminated what constitutes the teachers' L2 English performance and contribute to an ongoing debate about the features specific to second language production. They also add to an understanding of how dimensions of linguistic performance relate to each other and advance the CAF model of linguistic description of language production.

Study 2 has given an account of 132 students' test results at four different public elementary schools. 304 individual tests were administered, of which 49 students completed the BPVS3 at two test times and 87 students the ELIAS grammar test II at both test times. The aim was to examine how the students scored at each test time on the respective tests as well as how their scores developed over the fourth year between the two test times. The children's tests revealed manifold insights into the levels of attainment and development of two areas of foreign language learning at the primary level. As a whole group, the students did not statistically increase their receptive vocabulary as measured in the BPVS3 over their fourth year of elementary school, but showed a positive tendency. On average, the group of students significantly gained receptive grammar as measured in the ELIAS Grammar Test II. In addition, the findings revealed a great amount

of individual variance among the students. Individual variance has been reported repeatedly in research on children's development in the two areas of language.

Based on the interrelationships of the measures underlying each of the CAF dimensions, a Principal Component Analysis (PCA) of each dimension was computed in the synthesis of both previous studies in Study 3. In Part I, the large number of measures applied in the CAF analysis could be meaningfully reduced and conglomerated into the dimensions complexity, accuracy, speed fluency, breakdown fluency, repair fluency, and lexical diversity. Lexical diversity was not correlated with syntactic complexity measures – a finding that supports treating lexical diversity as its own dimension in future research and suggests extending the CAF framework to a CALF framework.

Syntactic complexity, accuracy, and breakdown fluency were significantly correlated. Language production hypotheses such as the Limited Capacity Hypothesis (1998, 2009) and the Cognition Hypothesis (Robinson, 2003, 2011), which assume the dimensions to come at the expense of one another, were not supported with the findings of the current studies. Instead, the results showed that the speakers drew on the three dimensions simultaneously, regardless of the relative language level among the group of teacher participants. There was no support for the assumption that complexity, accuracy, and fluency have to compete or that learners are always forced to choose between those competing dimensions in their production. However, the results cannot rule out that speakers may focus on one particular CAF dimension, for example if the tasks require them to do so or if the nature of the task is geared towards a stronger focus on one dimension. Speakers' performances may also change over time, which would only be observable in longitudinal with-in subject studies.

The findings in the synthesis study Part II, which answered how the teachers' L2 performance related to their students' receptive and grammar development, showed the specific nature of such a relationship. The relationships between the teachers' spoken performance and the students' outcomes of receptive grammar and vocabulary tests at two times reveal a multi-faceted picture. Based on existing research that found positive relationships between linguistic features in the input and children's target language acquisition, it was hypothesized that higher CAF performance of the teachers would correlate positively with the students' gain in receptive grammar and vocabulary. No significant positive relation could be found between comparably high speaking performances of teachers and higher tests scores of the students. To the contrary, there was a negative relationship between high-performance scores and the increase in the receptive grammar scores. If CAF performance is considered part of an overall target language proficiency, as was defined, the findings do not indicate that the more

proficient the teacher, the more their students will advance the L2 receptive vocabulary and grammar. Put differently, among the teachers examined in this study, the students of those teachers who scored comparably lower on language performance were nonetheless able to improve their receptive skills. Clearly, further studies are needed with more teachers who vary distinctly in their linguistic performance in order to support or refute the current findings. In addition, research that investigates the effects of linguistic performance in relation to other influencing factors on second language acquisition could illuminate the role of the teachers' performance as a covariate among those factors. Particularly, further studies on the relationship between over-all target language proficiency and beneficial teaching strategies could reveal more about the linkage between those two elements.

Novel insights have been offered in answering the question whether any particular dimension of the teachers' linguistic performance affected the students' receptive L2 development. Multiple regression analyses revealed that of all the dimensions applied in the CAF analyses, breakdown fluency was a significant predictor of the positive development in the students' receptive grammar. Whereas the overall picture did not show that teachers' higher performance came with a significant gain in their students' receptive vocabulary and grammar attainment, the opposite of which was hypothesized, the analysis of the individual CAF dimensions and their effects revealed a new finding. Breaking down speech in terms of pauses may have benefited the L2 receptive grammar development. This finding is in fact in line with some first as well as second language theoretical considerations: Pauses might help learners break the linguistic input into smaller chunks. Strings of language followed by pauses may be easier to process and give the learners time to disentangle meanings and forms of the foreign language input. The finding is also in accordance with teaching strategies that suggest modification in the input language to assist learners in taking in linguistic input, yet only with respect to pausing behavior. Pausing while speaking the L2 to learners of the L2 was expected to be beneficial.

Pausing may be particularly helpful to the acquisition of grammar, for which the segmentation of strings of language is essential. The children in the current studies were still at the beginning stage of learning English, at which the children's correct segmentation of the target language cannot be expected and still needs to be developed.

Several limitations underlie the current study. The interview Study 1 was limited to a one-shot interview design. Even though this was in accordance with the objective to measure the state of language performance, as opposed to development, several tasks might have been able to show a more diverse picture of the participants' language performance with regard to how they would perform on different tasks as well as at different times. Future research

may include a two-fold developmental approach that examines not only the students' L2 development, but also the teachers' language development.

Limitations are also given regarding the testing of the students' receptive grammar and vocabulary. Since the amount of English instruction was only two 45-minute lessons per week, the interval between the two times of testing can be considered relatively short to be able to show effects. A longer interval may be able to capture stronger time effects in particular on the vocabulary development, whose gain was not statistically significant in the current study. A specific limitation also surfaced considering some of the items in the standardized BPVS3: Some items are culturally and linguistically restricted to a UK variety of English, which may not be transferable to students growing up in a German environment and to L1 German-speaking English teachers. Future studies using a standardized vocabulary test may take this limitation into account and adapt some of the items.

Finally, a limitation in the synthesis Study 3 is the small number of teacher participants. Further studies including more teachers could illuminate similar or diverging findings on the relationship between teachers' L2 performance and students' second language development. In particular, further studies would be needed to examine the effects of specific CAF dimensions on the children's L2 development, such as breakdown fluency in the current study.

The present results have implications to several fields in second language acquisition. The variance in the L2 English teachers' language performance suggests that teachers' L2 language proficiency cannot be expected to be similar in a German elementary school context. This finding has implications for research on instructed second language acquisition in such a way that a highly proficient language level cannot be taken for granted, but that L2 English teachers rather vary a great deal. While this has also been acknowledged in studies that included teachers' proficiency as a factor, there is a need to take into account the potential of language providers' lower proficient language performance in models of second language acquisition and in theoretical considerations on beneficial language input. As was shown, models and theoretical debates on teaching strategies are implicitly or explicitly built on a perception of teachers as ideal model speakers of the L2 language. There is a need to adjust theoretical considerations to the great number of teachers who may not be considered ideal speakers of the L2. This appears to be crucial particularly in countries where the vast majority of English teachers at the primary level is not specialized in teaching English, as is the case for example in Germany.

The results regarding the children's development in receptive vocabulary and grammar suggest multiple implications as well. First, the students' significant mean gain in receptive grammar as a group and the positive,

albeit non-significant, tendency in receptive vocabulary over the course of the fourth grade imply that even at such a small amount of L2 instruction as two 45-minute lessons a week, L2 progress was noticeable. The positive development in the children's receptive vocabulary and grammar is a relevant result for the policy-making level, where decisions are made about if and when to introduce English as a foreign language in elementary schools. However, these findings need to be interpreted carefully: Whereas elementary children as a group may increase their receptive vocabulary and grammar skills, individual children may not do so. Thus, individual paths in language learning remain and need to be accounted for in any teaching approach.

Second, the analyses of the CAF dimensions and their relationships between one another indicate that second language speakers do not necessarily have to choose one dimension over another when producing spoken L2 language – a finding that may have implications for language training. If speakers can attend to complexity, accuracy, and fluency at the same time, language training can also focus on all the dimensions.

Third, the relationship between the breakdown fluency subdimension and the significant gain in grammar revealed some indications with respect to a specific feature in the teachers' L2 language performance that have implications for future studies incorporating the CAF framework. Investigating each dimension and subdimension individually may reveal more specific relationships that would otherwise stay unnoticed. The positive effect of the teachers' pausing on students' L2 development might be relief for L2 teachers, as breaking down one's stream of language by pausing may seem like an effortless language adjustment.

"The most precious things in speech are pauses", the bon mot by Sir Ralph Richardson leading into this thesis, may have come to an unexpected reality in second language acquisition. Yet the spaces between those pauses demand to be filled with language that is diverse in all aspects of language so as to offer a plentitude of language learning opportunities – a true challenge faced by the vast majority of L2 English teachers.

5 References

Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, (2), 1–47.

ACTFL Performance discriptors for language learners. (2012). Alexandria, USA. Retrieved from https://www.actfl.org/sites/default/files/pdfs/PerformanceDescriptorsLanguageLearners.pdf

ACTFL Proficiency guidelines. (2012). Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Ahmadian, M. J., & Tavakoli, P. (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), 35–59.

Aitchison, J. (2011). *The articulate mammal: An introduction to psycholinguistics* (5th ed.). New York: Routledge.

Alcañiz, E., & Muñoz, C. (2011). *The influence of parental and extramural factors on young learners of English. Unpublished Manuscript*. University of Barcelona.

Árva, V., & Medgyes, P. (2000). Native and non-native teachers in the classroom. *System*, *28*(3), 355–372. http://doi.org/http://doi.org/10.1016/S0346-251X(00)00017-8

Audacity (R): Free Audio Editor and Recorder [Computer program]. (2015). Audacity Team. Retrieved from https://www.audacityteam.org

Aukrust, V. G. (2007). Young children acquiring second language vocabulary in preschool group-time: Does amount, diversity, and discourse complexity of teacher talk matter? *Journal of Research in Childhood Education*, *22*(1), 17–37. http://doi.org/10.1080/02568540709594610

Bamanger, E. M., & Gashan, A. K. (2015). The effects of planning time on the fluency, accuracy, and complexity of EFL learners' oral production. *Journal of Educational Sciences*, *27*(1), 1–15.

Barnes, S., Gutfreund, M., & Satterly, D. (1983). Characteristics of adult speech which predict children's language development. *Journal of Child Language*, *10*, 65–84.

Bates, J. A. (2004). Use of narrative interviewing in everyday information behavior research. *Library & Information Science Research*, *26*(1), 15–28. http://doi.org/https://doi.org/10.1016/j.lisr.2003.11.003

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *13*(04), 525–531. http://doi.org/10.1017/s1366728909990423

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Bishop, D. V. M. (1989). *Test for reception of grammar (TROG)*. Manchester: Medical Research Council.

Bishop, D. V. M. (2003). *Test for reception of grammar – Version 2 (TROG–2)*. Ontario: Pearson Assessment.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2012). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *7*(30), 159–175. http://doi.org/10.1177/0265532212455394

Bowers, E. P., & Vasilyeva, M. (2011). The relation between teacher input and lexical growth of preschoolers. *Applied Psycholinguistics*, *32*, 221–241. http://doi.org/10.1017/S0142716410000354

Brandt, S., Kidd, E., Lieven, E., & Tomasello, M. (2009). The discourse bases of relativization: An investigation of young German and English-speaking children's comprehension of relative clauses. *Cognitive Linguistics*, *20*(3), 539–570. http://doi.org/10.1515/COGL.2009.024

Brockhaus, W. (2012). *Final devoicing in the phonology of German*. Berlin: De Gruyter.

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 21–46). Amsterdam/Philadelphia: John Benjamins.

Butler, Y. G. (2004). What level of English proficiency do elementary school teachers need to attain to teach EFL? Case studies from Korea, Taiwan, and Japan. *TESOL Quarterly*, *38*(2), 245. http://doi.org/10.2307/3588380

Buyl, A. (2010). *The development of receptive grammar knowledge in English as a second language: A cross-sectional study*. Master's Thesis. Vrije Universiteit Brussel. Retrieved from http://www.elias.bilikita.org/docs/Buyl_development_receptive_grammar.pdf.

Buyl, A., & Housen, A. (2015). Developmental stages in receptive grammar acquisition: A Processability Theory account. *Second Language Research*, *31*(4), 523–550. http://doi.org/10.1177/0267658315585905

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow: Longman.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873. http://doi.org/10.1016/j.cogsci.2003.06.001

Carroll, S. E. (2001). *Input and evidence: The raw material of second language acquisition*. Amsterdam/Philadelphia: John Benjamins. http://doi.org/10.1075/lald.25

Carroll, S. E. (2017). Exposure and input in bilingual development. *Bilingualism: Language and Cognition*, *20*(1), 3–16. http://doi.org/10.1017/S1366728915000863

Celce-Murcia, M., Brinton, D. M., & Snow, M. A. (Eds.). (2014). *Teaching English as a second or foreign language* (4th ed.). Boston, MA: National Geographic Learning.

Celce-Murcia, M., & McIntosh, L. (1991). *Teaching English as a second or foreign language*. New York, NY: Newbury House.

Chacón, C. T. (2005). Teachers' perceived efficacy among English as a foreign language teachers in middle schools in Venezuela. *Teaching and Teacher Education*, *21*(3), 257–272. http://doi.org/https://doi.org/10.1016/j.tate.2005.01.001

Chafe, W. (1988). Linking intonation units in spoken English. In J. Haiman & S. A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 1–27). Amsterdam: John Benjamins Publishing Company.

Chambless, K. S. (2012). Teachers' oral proficiency in the target language: Research on its role in language teaching and learning. *Foreign Language Annals*, *45*(s1), s141–s162. http://doi.org/10.1111/j.1944-9720.2012.01183.x

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton de Gruyter.

Chondrogianni, V., & Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguistic Approaches to Bilingualism*, *1*(3), 223–248. http://doi.org/10.1075/lab.1.3.05cho

Clark, E. (2009). *First language acquisition*. Cambridge: Cambridge University Press.

coat. (n.d.). In *Merriam-Webster.com*. Retrieved from https://www.merriam-webster.com/dictionary/coat [Retrieved March 8, 2018]

Cook, V. (2006). Basing teaching on the L2 user. In E. Llurda (Ed.), *Non-Native Language Teachers: Perceptions, Challenges and Contributions to the Profession* (pp. 47–61). Boston, MA: Springer US. http://doi.org/10.1007/0-387-24565-0_4

Cook, V. (2008). *Second language learning and language teaching* (4th ed.). London: Hodder Education.

Cook, V., & Singleton, D. (2014). *Key topics in second language acquisition*. Bristol, Buffalo, Toronto: Multilingual Matters.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. *Council of Europe*. http://doi.org/10.1017/S0267190514000221

Couve de Murville, S., Kersten, K., Maier, E., Ponto, K., & Weitz, M. (2016). Rezeptiver L2 Wortschatz in der Grundschule. In A. K. Steinlen & T. Piske (Eds.), *Wortschatzlernen in bilingualen Schulen und Kindertagesstätten* (pp. 85–121). Frankfurt a.M.: Peter Lang.

Crookes, G. (1989). Planning and interlanguage. *Studies in Second Language Acquisition*, *11*, 367–383. http://doi.org/https://doi.org/10.1017/S0272263100008391

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics*, *11*(2), 183–199. http://doi.org/10.1093/applin/11.2.183

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press. http://doi.org/10.1353/lan.2005.0220

Crystal, D. (2018a). *Sounds appealing: The passionate story of English pronunciation*. London: Profile Books.

Crystal, D. (2018b). *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press.

Csizér, K., & Dörnyei, Z. (2005). The internal structure of language learning motivation and its relationship with language choice and learning effort. *The Modern Language Journal*, *89*(1), 19–36.

Cullen, R. (2002). The use of lesson transcripts for developing teachers' classroom language. In H. R. Trappes-Lomax & G. Ferguson (Eds.), *Language in language teacher education* (pp. 219–235). Philadelphia: John Benjamins.

Davidson, C. (2009). Transcription: Imperatives for qualitative research. *International Journal of Qualitative Methods*, *8*(2), 1–52.

de Bot, K., Lowie, W., & Verspoor, M. (2007a). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7–21. http://doi.org/10.1017/S1366728906002732

de Bot, K., Lowie, W., & Verspoor, M. (2007b). A dynamic view as a complementary perspective. *Bilingualism: Language and Cognition*, *10*(1), 51–55. http://doi.org/10.1017/S1366728906002811

De Clercq, B., & Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research*, *35*(1), 71–97. http://doi.org/10.1177/0267658316674506

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *IRAL - International Review of Applied Linguistics in Language Teaching*, *54*(2), 113–132. http://doi.org/10.1515/iral-2016-9993

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)* (pp. 17–20). Stockholm: Royal Institute of Technology (KTH). Retrieved from http://hdl.handle.net/11858/00-001M-0000-0015-0FB8-8

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(02), 223–243. http://doi.org/10.1017/S0142716413000210

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893–916. http://doi.org/10.1017/S0142716412000069

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–90. http://doi.org/10.3758/BRM.41.2.385

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533–568. http://doi.org/10.1111/j.1467-9922.2010.00620.x

de Jong, N., & Vercellotti, M. Lou. (2015). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*(3), 387–404. http://doi.org/10.1177/1362168815606161

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533. http://doi.org/0272-2631/00

Dewaele, J.-M. (2009). Individual differences in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (2nd ed., pp. 623–646). Bingley, UK: Emerald Group Publishing.

Donzelli, G. (2007). Foreign language learners: Words they hear and words they learn. A case study. *Estudios de Lingüística Inglesa Aplicada*, (7), 103–125. http://doi.org/10.1021/nl400304y

Dörnyei, Z. (2005). *The psychology of the language learner. Individual differences in second language acquisition*. New York & London: Routledge Taylor & Francis Group.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. London & New York: Routledge.

Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research*. New York: Routledge. http://doi.org/https://doi.org/10.4324/9780203864739

Doughty, C. J., & Long, M. H. (Eds.). (2003). *The handbook of second language acquisition*. Malden/Oxford: Blackwell Publishing Ltd.

Dressler, R. A., & Kreuz, R. J. (2000). Transcribing oral discourse: A survey and a model system. *Discourse Processes*, *29*(1), 25–36. http://doi.org/10.1207/S15326950dp2901_2

Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. *Pragmatics*, *1*(1), 71–106.

Dunn, L. M., & Dunn, D. M. (2007a). *PPVT-4. Peabody Picture Vocabulary Test, Fourth Edition.* (4th ed.). Minneapolis: Pearson.

Dunn, L. M., & Dunn, D. M. (2007b). *PPVT-4 Manual*. Bloomington, MN: NCS Pearson, Inc.

Dunn, L. M., & Dunn, D. M. (2009). *The British Picture Vocabulary Scale BPVS: Third Edition. Manual*. London: GL Assessment.

Dunn, L. M., Dunn, D. M., Styles, B., & Sewell, J. (2009). *British Picture Vocabulary Scale BPVS: Third edition (BPVSIII)* (3rd ed.). London: GL Assessment.

Dunn, L. M., & Dunn, L. M. (1959). *Peobody Picture Vocabulary Test PPVT*. n.p.

Dunn, L. M., Dunn, L. M., Bulheller, S., & Häcker, H. (1965). *Peabody Picture Vocabulary Test PPVT*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., Dunn, L. M., & Whetton, C. (1982). *The British Picture Vocabulary Scale*. Windsor: NFER-Nelson.

Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). *The British Picture Vocabulary Scale* (2nd ed.). Windsor: NFER-Nelson.

Edwards, R., & Holland, J. (2013). *What is qualitative interviewing?* London, New Delhi, New York, Sydney: Bloomsbury Publishing.

Ellis, N. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(02), 143–188. http://doi.org/10.1017/S0272263102002024

Ellis, N., & Collins, L. (2009). Input and second language acquisition: The roles of frequency, form, and function. Introduction to the special issue. *Modern Language Journal*, *93*(3), 329–335. http://doi.org/10.1111/j.1540-4781.2009.00893.x

Ellis, N., & Wulff, S. (2015). Usage–based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 75–93). New York & London: Routledge.

Ellis, R. (1997). *Second language acquisition*. Oxford: OUP Oxford.

Ellis, R. (2005). Principles of instructed language learning. *System*, *33*(2), 209–224. http://doi.org/10.1017/CBO9781107415324.004

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474–509. http://doi.org/10.1093/applin/amp042

Ellis, R. (2015). *Understanding second language acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. London, New York: Taylor & Francis.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, *26*, 59–84. http://doi.org/10.1017/S0272263104261034

Enever, J. (Ed.). (2011). *ELLiE. Early language learning in Europe*. London: British Council. Retrieved from http://www.teachingenglish.org.uk/article/early-language-learning-europe

Enever, J. (2014). Primary English teacher education in Europe. *ELT Journal*, *68*(3), 231–242. http://doi.org/10.1093/elt/cct079

Eslami, Z. R., & Fatahi, A. (2008). Teachers' sense of self-efficacy, English proficiency, and instructional strategies: A study of nonnative EFL teachers in Iran. *Teaching English as a Second or Foreign Language - EJ*, *11*(4), 1–19. http://doi.org/10.1136/jmg.2004.027961

European Commission. (2005). A new framework strategy for multilingualism. *Communication from The Commission to The Council, The European Parliament, The European Economic And Social Committee and The Committee of the Regions*, 30. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0596:FIN:EN:PDF

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277–297). Amsterdam/ Philadelphia: John Benjamins.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Los Angeles, London, New Delhi, Singapore, Washington D.C.: SAGE Publications.

fit. (n.d.). In *dictionarycambridge.org*. Retrieved from https://dictionary.cambridge.org/grammar/british-grammar/types-of-english-formal-informal-etc/british-and-american-english [accessed Feb. 27, 2019]

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

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, *59*(4), 866–896. http://doi.org/10.1111/j.1467-9922.2009.00528.x

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. http://doi.org/DOI: 10.1093/applin/21.3.354

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a Weighted Clause Ratio. *Annual Review of Applied Linguistics*, *36*, 98–116. http://doi.org/10.1017/S0267190515000082

Gardner, R. C., Masgoret, A.-M., & Tremblay, P. F. (1999). Home background characteristics and second language learning. *Journal of Language and Social Psychology*, *18*(4), 419–437.

Gass, S. M. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, *9*(2), 198–217. http://doi.org/https://doi.org/10.1093/applin/9.2.198

Gass, S. M. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S. M. (2003). Input and interaction. In M. H. Long & C. Doughty (Eds.), *The handbook of second language acquisition* (pp. 224–255). Blackwell Publishing Ltd.

Gass, S. M. (2013). *Second language acquisition: An introductory course* (4th ed.). New York: Routledge.

Gass, S. M. (2018). *Input, interaction, and the second language learner*. New York & London: Routledge Taylor & Francis Group.

Gass, S. M., & Selinker, L. (Eds.). (2001). *Second language acquisition. An introductory course* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition. An introductory course* (3rd ed.). New York & London: Routledge Taylor & Francis Group.

Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, *16*, 283–302.

Ghasemboland, F., & Hashim, F. B. (2013). Teachers' self-efficacy beliefs and their English language proficiency: A study of nonnative EFL teachers in selected language centers. *Procedia - Social and Behavioral Sciences*, *103*, 890–899. http://doi.org/https://doi.org/10.1016/j.sbspro.2013.10.411

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and [+/-Here-and-Now]: Effects on L2 oral production. In M. Garcia-Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 44–68). Clevedon: Multilingual Matters.

Gnewuch, N. (2014). *Testing receptive vocabulary knowledge in second language acquisition*. Unpublished Master's Thesis. University of Hildesheim.

Golberg, H., Paradis, J., & Crago, M. (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Applied Psycholinguistics*, *29*(1), 41–65. http://doi.org/10.1017/S014271640808003X

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

Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2017). Early language learning: The impact of teaching and teacher factors. *Language Learning*, *67*(4), 922–958. http://doi.org/10.1111/lang.12251

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Studies in syntax and semantics III: Speech acts* (pp. 41–58). New York: Academic Press.

Gumperz, J. J., & Berenz, N. (1993). Transcribing conversational exchanges. In J. A. Edwards & M. D. Lampert (Eds.), *Transcription and coding methods for language research* (pp. 91–122). Hillsdale, N.J.: Lawrence Erlbaum.

Harley, B., Allen, P., Cummins, J., & Swain, M. (Eds.). (1990). *The development of second language proficiency*. Cambridge: Cambridge University Press.

Harmer, J. (2007). *The practice of English language teaching* (4th ed.). Harlow: Pearson Longman.

Harris, J. (2002). *Early language development: Implications for clinical and educational practice*. London, New York: Taylor & Francis.

Hatch, E. M. (1983). *Psycholinguistics: A second language perspective*. Rowley, MA: Newbury House Publishers.

Hoff-Ginsberg, E. (1985). Some contributions of mothers' speech to their children's syntactic growth. *Journal of Child Language*, *12*, 367–385.

Hoff, E. (2003). Causes and consequences of SES-related differences in parent-to-child speech. In M. H. Bornstein & R. H. Bradley (Eds.), *Monographs in parenting series. Socioeconomic status, parenting, and child development* (pp. 147–160). Mahwah, NJ: Lawrence Erlbaum Associates.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88. http://doi.org/https://doi.org/10.1016/j.dr.2005.11.002

Holmes, J., & Wilson, N. (2017). *An introduction to sociolinguistics* (5th ed.). London & New York: Routledge.

Hopp, H., Kieseier, T., Vogelbacher, M., & Thoma, D. (2018). L1 effects in the early L3 acquisition of vocabulary and grammar. In A. Bonnet & P. Siemund (Eds.), *Foreign language education in multilingual classrooms* (pp. 305–330). Amsterdam: John Benjamins.

Horváth, J., & Nikolov, M. (Eds.). (2007). *UPRT 2007: Empirical studies in English applied linguistics*. Pécs: Lingua Franca Csoport. http://doi.org/10.1017/CBO9781107415324.004

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. http://doi.org/10.1093/applin/amp048

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

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012b). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. Amsterdam/Philadelphia: John Benjamins.

Huang, B. H., Chang, Y. H. S., Zhi, M., & Niu, L. (2018). The effect of input on bilingual adolescents' long-term language outcomes in a foreign language instruction context. *International Journal of Bilingualism*, *00*(0), 1–18. http://doi.org/10.1177/1367006918768311

Hudson, R. A. (1996). *Sociolinguistics* (2nd ed.). Cambridge: Cambridge University Press.

Hughes, A., Trudgill, P., & Watt, D. (2013). English accents and dialects: An introduction to social and regional varieties of English in the British Isles. London & New York: Routledge.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels. NCTE Research Report No. 3.* Washington, DC: ERIC.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*, 337–374.

IELTS (International English Language Testing System). (2007). *Handbook 2007*. www.ielts.org.

Inbar-Lourie, O. (2010). English only? The linguistic choices of teachers of young EFL learners. *International Journal of Bilingualism*, *14*(3), 351–367.

Information and Privacy Commissioner of Ontario. (2015). *Best practices for protecting individual privacy in conducting survey research*. Ontario Public Service. Retrieved from https://www.ipc.on.ca/wp-content/uploads/2015/04/best-practices-for-protecting-individual-privacy-in-conducting-survey-research.pdf

Inoue, C. (2010). Investigating the sensitivity of the measures of fluency, accuracy, complexity and idea units with a narrative task. *Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching*, *4*, 1–24.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. http://doi.org/10.1093/applin/amm017

jacket. (n.d.). In *Merriam-Webster.com*. Retrieved from https://www.merriam-webster.com/dictionary/jacket [Retrieved March 8, 2018]

Jaekel, N., Schurig, M., & Florian, M. (2017). From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning*, *xxxx*, 1–34. http://doi.org/10.1111/lang.12242

Jarvis, S., & Daller, M. (Eds.). (2013). *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam/Philadelphia: John Benjamins.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 13–31). Amsterdam/Philadelphia: John Benjamins.

Jenkins, J. (2015). *Global Englishes* (3rd ed.). London & New York: Routledge.

Johnson, D. M. (1992). *Approaches to research in second language learning*. New York: Longman.

Johnson, R., Onwuegbuzie, A., & Turner, L. (2007). Towards a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2), 112–138. http://doi.org/10.1017/9781316418376.015

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer. http://doi.org/10.2307/1270093

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21. http://doi.org/10.1016/j.cogpsych.2017.07.002

Kawaguchi, S. (2013). The relationship between lexical and syntactic development in English as a second language. In A. Flyman Mattsson & C. Norrby (Eds.), *Language acquisition and use in multilingual contexts* (pp. 92–106). Lund: Travaux de l'Institut de Linguistique de Lund; Vol. 52.

Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 143–164). Amsterdam/Philadelphia: John Benjamins.

Kersten, K. (2019). Einflussfaktoren im bilingualen Fremdsprachenerwerb. In A. Rohde & A. K. Steinlen (Eds.), *Spachenvielfalt als Ressource begreifen. Mehrsprachigkeit in bilingualen Kindertagesstätten und Schulen. Band II* (pp. 35–70). Berlin: dohrmann Verlag.

Kersten, K., Frey, E., & Hähnert, A. (2008). *ELIAS Early Language and Intercultural Acquisition Studies*. Retrieved from www.fmks-online.de/_wd_showdoc.php?pic=690

Kersten, K., Piske, T., Rohde, A., Steinlen, A. K., Weitz, M., & Kurth, S. (2010). *ELIAS Grammar Test*. Magdeburg: Universität Magdeburg ELIAS. Retrieved from www.bilikita.org

Kersten, K., Piske, T., Rohde, A., Steinlen, A., Weitz, M., & Couve de Murville, S. (n.d.). *ELIAS Grammar Test II. Unpublished test*.

Kersten, K., Schüle, C., & Steinlen, A. K. (forthc.). Variables affecting early foreign language acquisition (Preprint), *submitted*.

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

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality and aptitude in narrative task performance. *Language Learning*, *62*(2), 439–472.

Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford: Oxford University Press.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. New York: Longman Ltd.

Krashen, S. (2009). The comprehension hypothesis extended. In T. Piske & M. Young-Scholten (Eds.), *Input matters in SLA* (pp. 81–94). Bristol, Buffalo, Toronto: Multilingual Matters. http://doi.org/10.1017/CBO9781107415324.004

Kuiken, F., Vedder, I., Housen, A., & De Clerq, B. (Eds.). (2019). Special Issue on Syntactic Complexity [Special Issue]. *International Journal of Applied Linguistics*, *29*(2), 159–282.

Kultusmisterkonferenz KMK. (2013). Bericht „Fremdsprachen in der Grundschule – Sachstand und Konzeptionen 2013". Retrieved from http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2013/2013_10_17-Fremdsprachen-in-der-Grundschule.pdf

Kuppens, A. H. (2010). Incidental foreign language acquisition from media exposure. *Learning, Media and Technology*, *35*(1), 65–85. http://doi.org/10.1080/17439880903561876

Lambert, C., & 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. http://doi.org/10.1093/applin/amu047

Lapadat, J. C., & Lindsay, A. C. (1999). Transcription in research and practice: From standardization of technique to interpretive positionings. *Qualitative Inquiry*, *5*(1), 64–86.

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. http://doi.org/10.1093/applin/aml029

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579–589. http://doi.org/10.1093/applin/amp043

Larsen-Freeman, D. (2015). Complexity theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 227–244). New York & London: Routledge.

Larsen-Freeman, D., & Long, M. H. (1991). *An introduction to second language acquisition research*. London & New York: Routledge.

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, *24*(1), 35–63. http://doi.org/https://doi.org/10.1177/0267658307082981

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R.* (2nd ed.). New York: Routledge. http://doi.org/https://doi.org/10.4324/9781315775661

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. http://doi.org/10.1093/applin/22.1.1

Leclercq, P., Edmonds, A., & Hilton, H. (Eds.). (2014). *Measuring L2 proficiency: Perspectives from SLA*. Bristol, Buffalo, Toronto: Multilingual Matters.

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

Leow, R. P. (1993). To simplify or not to simplify: A look at intake. *Studies in Second Language Acquisition*, *15*(3), 333–355. http://doi.org/10.1017/S0272263100012146

Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York & London: Routledge Taylor & Francis Group.

Leow, R. P. (2019). Theoretical underpinnings and cognitive processes in instructed SLA. In R. P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning: Processing and processes*. New York & London: Taylor & Francis.

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

Lieven, E. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 56–73). Cambridge: Cambridge University Press. http://doi.org/DOI: 10.1017/CBO9780511620690.005

Lieven, E. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, *120*, 2546–2556. http://doi.org/10.1016/j.lingua.2010.06.005

Lightbown, P. M. (2014). Making the minutes count in L2 teaching. *Language Awareness*, *23*(1–2), 3–23. http://doi.org/10.1080/09658416.2013.863903

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, *10*(1), 105–129. http://doi.org/10.1080/14790718.2012.679275

Lintunen, P., & Mäkilä, M. (2015). Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in Language*, *12*(4), 377–399. http://doi.org/10.1515/rela-2015-0005

Little, T. D. (Ed.). (2013). *The Oxford handbook of quantitative methods. Volume 2: Statistical analysis*. New York: Oxford University Press.

Llurda, E. (2004). Non-native-speaker teachers and English as an international language. *International Journal of Applied Linguistics*, *14*(3), 314–323. http://doi.org/10.1111/j.1473-4192.2004.00068.x

Llurda, E. (Ed.). (2006). *Non-native language teachers: Perceptions, challenges and contributions to the profession*. New York: Springer Science & Business Media.

Loban, W. (1966). *Problems in oral English: Kindergarten through grade nine*. Champaign, Ill. Retrieved from https://eric.ed.gov/?id=ED070106

Loder Buechel, L. (2015). *Associations between young learners' English language performance and teacher proficiency and experience with English. Dissertation*. Dissertation. University of Fribourg, Switzerland. Retrieved from https://core.ac.uk/download/pdf/43662001.pdf

Loewen, S., & Sato, M. (2018). Interaction and instructed second language acquisition. *Language Teaching*, *51*(3), 285–329. http://doi.org/10.1017/s0261444818000125

Long, M. H. (1981). Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences*, *379*(1), 259–278. http://doi.org/10.1111/j.1749-6632.1981.tb42014.x

Lowie, W., Verspoor, M., & van Dijk, M. (2018). The acquisition of L2 speaking. In R. A. Alonso (Ed.), *Speaking in a Second Language* (pp. 105–126). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Lüke, T., Ritterfeld, U., & Tröster, H. (2016). Erprobung eines Gruppentests zur Überprüfung des Grammatikverständnisses auf der Basis des TROG-D. *Diagnostica*, *62*, 242–254. http://doi.org/https://doi.org/10.1026/0012-1924/a000157

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Mackey, A., & Gass, S. M. (2015). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition. An introduction* (2nd ed., pp. 180–206). New York & London: Routledge.

Mackey, A., & Gass, S. M. (2016). *Second language research. Methodology and design* (2nd ed.). New York & London: Routledge.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–295. http://doi.org/10.1017/S0305000900006449

Maier, E., Neubauer, L., Ponto, K., Couve de Murville, S., & Kersten, K. (2016). Assessing linguistic levels of L2 English in primary school programs. In J.-U. Keßler, A. Lenzing, & M. Liebner (Eds.), *Developing, modelling and assessing second languages* (pp. 163–192). John Benjamins. http://doi.org/10.1075/palart.5.08mai

Malvern, D., & Richards, B. (2002). Investigating accommodation in language profciency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85–104.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). Lexical diversity and the investigation of accommodation in foreign language proficiency interviews. In D. Malvern, B. Richards, N. Chipere, & P. Durán (Eds.), *Lexical diversity and language development: Quantification and assessment* (pp. 95–109). London: Palgrave Macmillan UK. http://doi.org/10.1057/9780230511804_6

Mani, N., & Pätzold, W. (2016). Sixteen-month-old infants segment words from infant- and adult-directed speech. *Language Learning and Development*, *12*(4), 499–508. http://doi.org/10.1080/15475441.2016.1171717

Matsuda, S., & Gobel, P. (2004). Anxiety and predictors of performance in the foreign language classroom. *System*, *32*(1), 21–36. http://doi.org/https://doi.org/10.1016/j.system.2003.08.002

McCarthy, P. M. (2017). Comments on Text Inspector. Retrieved from https://textinspector.com/help/lexical-diversity/

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. http://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. http://doi.org/10.3758/BRM.42.2.381

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*(3), 323–338. http://doi.org/https://doi.org/10.1093/llc/15.3.323

Medgyes, P. (1992). Native or non-native: who's worth more? *ELT Journal*, *46*(4), 340–349.

Medgyes, P. (2001). When the teacher is a non-native speaker. *Teaching English as a Second or Foreign Language*. http://doi.org/10.1093/elt/ccq092

Media Consulting Group. (2008). *Study on dubbing and subtitling needs and practices in the European audiovisual industry*. Retrieved from https://www.lt-innovate.org/sites/default/files/documents/1342-Study on dubbing and subtitling needs and practices in the European audiovisual industry%282008%29.pdf

Meisel, J. (1977). *Languages in contact*. Tübingen: Narr.

Meisel, J. (2011). *First and second language acquisition. Parallels and differences*. Cambridge: Cambridge University Press.

Mesthrie, R., & Bhatt, R. M. (2008). *World Englishes. The study of new linguistic varieties (Key Topics in Sociolinguistics)*. Cambridge: Cambridge University Press.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics*, *45*, 241–59.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, Buffalo, Toronto: Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. *EUROSLA Monographs Series 2. L2 Vocabulary Acquisition, Knowledge and Use*, 57–78.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117. http://doi.org/10.1016/S0010-0277(03)00140-9

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *Tesol Quarterly*, *46*(4), 610–641.

Muñoz, C. (2008). Age-related differences in foreign language learning. Revisiting the empirical evidence. *IRAL - International Review of Applied Linguistics in Language Teaching*, *46*(3), 197–220. http://doi.org/10.1515/IRAL.2008.009

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, *35*(4), 463–482. http://doi.org/10.1093/applin/amu024

Myles, F., Mitchell, R., & Marsden, E. (2013). *Second language learning theories* (3nd ed.). London, New York: Routledge.

Nel, N., & Müller, H. (2010). The impact of teachers' limited English proficiency on English second language learners in South African schools. *South African Journal of Education*, *30*, 635–650.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children* (pp. 109–149). Cambridge: Cambridge University Press.

Newton, J. (2013). Incidental vocabulary learning in classroom communication tasks. *Language Teaching Research*, *17*(2), 164–187. http://doi.org/10.1177/1362168812460814

Niedersächsisches Kultusministerium. (2006). *Kerncurriculum für die Grundschule Schuljahrgänge 3–4 Englisch*. Hannover. Retrieved from http://db2.nibis.de/1db/cuvo/ausgabe/

Niedersächsisches Kultusministerium. (2018). *Kerncurriculum für die Grundschule Schuljahrgänge 3-4 Englisch*. Hannover. Retrieved from http://www.cuvo.nibis.de

Nikolov, M., & Mihaljević Djigunović, J. (2011). All shades of every color: An overview of early teaching and learning of foreign languages. *Annual Review of Applied Linguistics*, *31*, 95–119. http://doi.org/DOI: 10.1017/S0267190511000183

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4f), 555–578. http://doi.org/10.1093/applin/amp044

Nunan, D. (1999). *Second language teaching & learning*. Florence, KY: Heinle & Heinle Publishers.

Nunan, D. (2015). *Teaching English to speakers of other languages*. New York: Routledge. http://doi.org/doi.org/10.4324/9781315740553

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading*, *16*(2), 91–121. http://doi.org/10.1080/10888438.2010.529219

Oliver, R., & Azkarai, A. (2017). Review of child second language acquisition (SLA): Examining theories and research. *Annual Review of Applied Linguistics*, *37*(2017), 62–76. http://doi.org/10.1017/S0267190517000058

Ortega, L. (2009). *Understanding second language acquisition*. London & New York: Routledge.

Ortega, L. (Ed.). (2011). *Second language acquisition. Critical concepts in linguistics* (Vol. VI). London & New York: Routledge.

Ortega, L. (2015). Second language learning explained? SLA across 10 contemporary theories. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 245–272). New York & London: Routledge.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. http://doi.org/10.1093/applin/amp045

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117–134. http://doi.org/10.1177/0267658314536435

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, *76*(4), 763–782.

Paradis, J. (2011). Individual differences in child English second language acquisition. *Linguistic Approaches to Bilingualism*, *1*(3), 213–237. http://doi.org/10.1075/lab.1.3.01par

Paradis, J., Nicoladis, E., Crago, M., & Genesee, F. (2011). Bilingual children's acquisition of the past tense: A usage-based approach. *Journal of Child Language*, *38*(3), 554–578. http://doi.org/10.1017/S0305000910000218

Paradis, J., Rusk, B., Sorenson Duncan, T., & Govindarajan, K. (2017). Children's second language acquisition of English complex syntax: The role of age, input, and cognitive factors. *Annual Review of Applied Linguistics*, *37*, 148–167. http://doi.org/10.1017/S0267190517000022

Paulus, T., Lester, J., & Dempster, P. (2014). *Digital tools for qualitative research*. London: SAGE Publications.

Pearson, B. Z., Fernandez, S. C., Lewedeg, V., & Oller, K. D. (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics*, *18*(1), 41–58. http://doi.org/10.1017/S0142716400009863

Pedersen, J. (2011). *Subtitling norms for television: An exploration focussing on extralinguistic cultural references*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pica, T. (2005). Second language acquisition research and applied linguistics (Postprint version). *University of Pennsylvania Scholarly Commons*, *1*(1). Retrieved from http://repository.upenn.edu/gse_pubs/34

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

Pine, J. M. (1994). The language of primary caregivers. In C. Gallaway & B. J. E. Richards (Eds.), *Input and interaction in language acquisition* (pp. 15–37). Cambridge University Press. http://doi.org/10.1017/CBO9780511620690.003

Pinget, A.-F., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31*(3), 349–365. http://doi.org/10.1177/0265532214526177

Pinker, S. (2004). Clarifying the logical problem of language acquisition. *Journal of Child Language*, *31*(04), 949–953. http://doi.org/10.1017/S0305000904006439

Piske, T. (2011, December 6). Fremdsprachen für Grundschüler: "Früh anfangen allein reicht nicht." *Spiegel Online*. Retrieved from http://www.spiegel.de/lebenundlernen/schule/fremdsprachen-fuer-grundschueler-frueh-anfangen-allein-reicht-nicht-a-788653.html

Rankin, T., & Unsworth, S. (2016). Beyond poverty: Engaging with input in generative SLA. *Second Language Research*, *32*(4), 563–572. http://doi.org/10.1177/0267658316648732

Révész, A., Ekiert, M., & Torgersen, E. N. (2014a). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. Supporting information 2nd revision. *Applied Linguistics*. Retrieved from http://applij.oxfordjournals.org/content/early/2014/12/03/applin.amu069.abstract

Révész, A., Ekiert, M., & Torgersen, E. N. (2014b). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 1–22. http://doi.org/10.1093/applin/amu069

Richards, B. J. (1994). Child-directed speech and influences on language acquisition: Methodology and interpretation. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 74–106). New York: Cambridge University Press.

Richards, H. M., Conway, C., Roskvist, A., Harvey, S., Richards, H. M., Conway, C., … Harvey, S. (2012). Foreign language teachers' language proficiency and their language teaching practice. *Language Learning Journal*, *41*(2), 231–246. http://doi.org/10.1080/09571736.2012.707676

Richards, J. C. (2017). Teaching English through English: Proficiency, pedagogy and performance. *RELC Journal*, *48*(1), 7–30. http://doi.org/10.1177/0033688217690059

Richards, J. C., & Lockhart, C. (1994). *Reflective teaching in second language classrooms*. Cambridge: Cambridge University Press.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching* (3rd ed.). Cambridge, UK: Cambridge University Press.

Richards, J. C., & Schmidt, R. (2010). *Longman dictionary of language teaching and applied linguistics* (4th ed.). Harlow: Pearson Education.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423–441. http://doi.org/10.1080/01638539109544795

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, (45), 99–140.

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 631–678). Malden, MA: Blackwell Publishing Ltd. http://doi.org/10.1002/9780470756492.ch19

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–38). Amsterdam/ Philadelphia: John Benjamins.

Rohde, A. (2010). Receptive L2 lexical knowledge in bilingual preschool children. In K. Kersten, A. Rohde, C. Schelletter, & A. K. Steinlen (Eds.), *Bilingual preschools. Volume I. Learning and development* (pp. 45–68). Trier: WVT Wissenschaftlicher Verlag Trier.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(1), 185–205. http://doi.org/10.1017/S0305000907008343

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech vocabulary development. *Child Development*, *83*(5), 1762–1774. http://doi.org/10.1111/j.1467-8624.2012.01805.x

Ryan, B. P. (1992). Articulation, language, rate, and fluency characteristics of stuttering and nonstuttering preschool children. *Journal of Speech, Language, and Hearing Research*, *35*(2), 333–342.

Saito, K., & Hanzawa, K. (2018). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*, *22*(4), 398–417. http://doi.org/10.1177/1362168816679030

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly*, *50*(2), 421–446. http://doi.org/10.1002/tesq.234

Sample, E., & Michel, M. (2014). An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. *TESL Canada Journal*, *31*(8), 23–46.

Saville-Troike, M. (2016). *Introducing second language acquisition* (3rd ed.). Cambridge: Cambridge University Press.

Schelletter, C., & Ramsey, R. (2010). Lexical and grammatical comprehension in monolingual and bilingual children. In K. Kersten, A. Rohde, C. Schelletter, & A. K. Steinlen (Eds.), *Bilingual preschools. Volume 1. Learning and development.* (pp. 101–118). Trier: WVT Wissenschaftlicher Verlag. Retrieved from https://www.researchgate.net/profile/Christina_Schelletter/publication/233025555_Lexical_and_Grammatical_Comprehension_in_Monolingual_and_Bilingual_Children/links/0912f509b9e292c34b000000.pdf

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. http://doi.org/10.1093/applin/11.2.129

Schmidt, R. (1994a). Deconstructing consciousness in search of useful definitions for applied linguistics. *Consciousness in Second Language Learning*, *11*, 237–326.

Schmidt, R. (1994b). Deconstructing consciousness in search of useful definitions for applied linguistics. In J. H. Hulstijn & R. Schmidt (Eds.), *Consciousness in second language learning* (Vol. 11, pp. 11–26). AILA Review.

Schreiner, M. S., & Mani, N. (2017). Listen up! Developmental differences in the impact of IDS on speech segmentation. *Cognition*, *160*, 98–102. http://doi.org/https://doi.org/10.1016/j.cognition.2016.12.003

Seedhouse, P., Harris, A., Naeb, R., & Üstünel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Report Series*, (2), 1–30. Retrieved from www.ielts.org/researchers

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

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 79–95. http://doi.org/10.1515/iral-2016-9991

Seidlhofer, B. (2013). *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Sharwood Smith, M. (1994). *Second language learning: Theoretical foundations*. London & New York: Routledge.

Silverman, S., & Ratner, N. B. (2002). Measuring lexical diversity in children who stutter: Application of vocd. *Journal of Fluency Disorders*, *27*(4), 289–304. http://doi.org/10.1016/S0094-730X(02)00162-6

Skapinker, M. (2007, November 8). Whose language? *Financial Times*. Retrieved from https://www.ft.com/content/e621ff38-8e1c-11dc-8591-0000779fd2ac

Skehan, P. (1989). *Individual differences in second-language learning*. London: Edward Arnold.

Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition*, *13*, 275–298. http://doi.org/doi.org/10.1017/S0272263100009979

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

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*, 1–14. http://doi.org/https://doi.org/10.1017/S026144480200188X

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. http://doi.org/10.1093/applin/amp047

Skehan, P. (2015). Limited attention capacity and cognition. Two hypotheses regarding second language performance on tasks. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference* (pp. 123–156). Amsterdam: John Benjamins. http://doi.org/10.1075/tblt.8.05ske

Skehan, P. (2018). *Second language task-based performance*. New York & London: Routledge Taylor & Francis Group.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*, 185–211.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *IRAL - International Review of Applied Linguistics in Language Teaching*, *54*(2), 97–111. http://doi.org/10.1515/iral-2016-9992

Smithson, L., Paradis, J., & Nicoladis, E. (2014). Bilingualism and receptive vocabulary achievement: Could sociocultural context make a difference? *Bilingualism: Language and Cognition*, *17*(4), 810–821. http://doi.org/10.1017/S1366728913000813

Snow, C. E. (1995). Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 179–193). Oxford: Blackwell Publishing Ltd.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. http://doi.org/https://doi.org/10.1016/j.dr.2007.06.002

Sorenson Duncan, T., & Paradis, J. (2018). How does maternal education influence the linguistic environment supporting bilingual language development in child L2 learners of English? *International Journal of Bilingualism*, 1–16. http://doi.org/doi.1367006918768366

Statista. (2019). Smartphone-Besitz bei Kindern und Jugendlichen in Deutschland im Jahr 2017 nach Altersgruppe. Retrieved from https://de.statista.com/statistik/daten/studie/1106/umfrage/handybesitz-bei-jugendlichen-nach-altersgruppen/

Steinlen, A. K. (2013). "Flera språk–fler möjligheter"–Immigrant children's acquisition of English in bilingual preschools. In A. Flyman Mattsson & C. Norrby (Eds.), *Language acquisition and use in multilingual contexts* (pp. 170–184). Lund: Lund University.

Steinlen, A. K., Håkansson, G., Housen, A., & Schelletter, C. (2010). Receptive L2 grammar knowledge development in bilingual preschools. In K. Kersten, A. Rohde, C. Schelletter, & A. K. Steinlen (Eds.), *Bilingual preschools, Volume I. Learning and development* (pp. 69–100). Trier: WVT Wissenschaftlicher Verlag.

Steinlen, A. K., & Piske, T. (Eds.). (2016). *Bilinguale Programme in Kindertageseinrichtungen: Umsetzungsbeispiele und Forschungsergebnisse*. Tübingen: Narr.

Steinlen, A. K., & Rogotzki, N. (2008). Comprehension of L2 grammar in a bilingual preschool: A developmental perspective. In *Proceedings of the child language seminar 2007* (pp. 163–173). Reading: University of Reading. Retrieved from http://www.reading.ac.uk/web/files/cls/CLS_Steinlen,Rogotzki.pdf

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 133–150. http://doi.org/10.1515/iral-2016-9994

Tavakoli, P. (2018). L2 development in an intensive Study Abroad EAP context. *System*, *72*, 62–74. http://doi.org/10.1016/j.system.2017.10.009

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447–471. http://doi.org/10.1002/tesq.244

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam/Philadelphia: John Benjamins.

Text Inspector. (2016). Online lexis analysis tool at textinspector.com. Retrieved from https://textinspector.com/ [Accessed 2016]

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71. http://doi.org/10.1207/s15327078in0701_5

Thornbury, S. (2000). Accuracy, fluency and complexity. *English Teaching Professional*, *16 July*, 3–6.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84–119. http://doi.org/10.1093/applin/17.1.84

Traxler, M., & Gernsbacher, M. A. (Eds.). (2006). *Handbook of psycholinguistics* (2nd ed.). Amsterdam et al.: Academic Press.

Trebits, A., Adler, M., & Kersten, K. (forthc.). Cognitive variables and socioeconomic status in early second language acquisition, *forthc.*

Trebits, A., & Kersten, K. (forthc.). Bilingual education trumps socioeconomic status: L1 and L2 development in primary school learners of English as a foreign language, *forthc.*

Trudgill, P. (2000). *Sociolinguistics: An introduction to language and society* (4th ed.). London: Penguin UK.

Trudgill, P., & Hannah, J. (2013). *International English: A guide to the varieties of standard English*. London & New York: Routledge.

Unsworth, S. (2016a). Early child L2 acquisition: Age or input effects? Neither, or both? *Journal of Child Language*, *43*(3), 608–634. http://doi.org/10.1017/S030500091500080X

Unsworth, S. (2016b). Quantity and quality of language input in bilingual language development. In E. Nicoladis & S. Montanari (Eds.), *Lifespan perspectives on bilingualism: Factors moderating language proficiency* (pp. 136–196). Washington, DC & Berlin: American Psychological Association & Walter de Gruyter. Retrieved from http://sharonunsworth.org/Publications_files/Bilingualism_lifespan_FINAL.pdf

Unsworth, S., Persson, L., Prins, T., & Bot, K. de. (2015). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*, *36*(5), 527–548. http://doi.org/10.1093/applin/amt052
Urdan, T. C. (2017). *Statistics in plain English* (4th ed.). New York & London: Taylor & Francis.

Van Canh, L., & Renandya, W. A. (2017). Teachers' English proficiency and classroom language use: A conversation analysis study. *RELC Journal*, *48*(1), 67–81. http://doi.org/10.1177/0033688217690935
VanPatten, B. (1990). Attending to form. An experiment in consciousness. *Studies in Second Language Acquisition*, *12*, 287–301. http://doi.org/https://doi.org/10.1017/S0272263100009177
VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex Publishing.
VanPatten, B., & Benati, A. G. (2015). *Key terms in second language acquisition*. London & New York: Bloomsbury Publishing.
Vercellotti, M. Lou. (2012). *Complexity, accuracy, and fluency as properties of language performance: The development of the multiple subsystems over time and in relation to each other*. PhD Thesis. University of Pittsburgh. http://doi.org/10.1017/CBO9781107415324.004
Vercellotti, M. Lou. (2015). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 1–23. http://doi.org/10.1093/applin/amv002
Vercellotti, M. Lou. (2019). Finding variation: assessing the development of syntactic complexity in ESL Speech. *International Journal of Applied Linguistics*, *29*(2), 233–247. http://doi.org/10.1111/ijal.12225
Verspoor, M., Lowie, W., Chan, H. P., & Vahtrick, L. (2017). Linguistic complexity in second language development: variability and variation at advanced stages. *Recherches En Didactique Des Langues et Des Cultures*, *14*(1), 0–28. http://doi.org/10.4000/rdlc.1450

waistcoat. (n.d.). In *Merriam-Webster.com*. Retrieved from https://www.merriam-webster.com/words-at-play/words-commonly-mispronounced/waistcoat [accessed Feb. 13, 2017]
Wang, L. L., Watts, A. S., Anderson, R. A., & Little, T. D. (2013). Common fallacies in quantitative research methodology. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology. Vol. 2: Statistical analysis* (pp. 718–758). Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199934898.013.0031
Warren, P. (2016). *Uptalk: The phenomenon of rising intonation*. Cambridge: Cambridge University Press.
Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95.

Weitz, M. (2015). *Die Rolle des L2-Inputs in bilingualen Kindergärten*. Frankfurt a.M.: Peter Lang.

Weitz, M., Pahl, S., Flyman Mattsson, A., Buyl, A., & Kalbe, E. (2010). The Input Quality Observation Scheme (IQOS): The nature of L2 input and its influence on L2 development in bilingual preschools. In K. Kersten, A. Rohde, C. Schelletter, & A. K. Steinlen (Eds.), *Bilingual preschools, Volume I. Learning and development* (pp. 5–44). Trier: Wissenschaftlicher Verlag Trier.

Wesche, M. B. (1994). Input and interaction in second language acquisition. In C. Gallaway & J. B. Richards (Eds.), *Input and interaction in language acquisition* (pp. 219–249). Cambridge: Cambridge University Press.

White, L. (1987). Against Comprehensible Input: The Input Hypothesis and the development of second-language competence. *Applied Linguistics*, *8*(2), 95–110. http://doi.org/doi.org/10.1093/applin/8.2.95

White, L. (2015). Linguistic theory, universal grammar, and second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 34–53). New York & London: Routledge Taylor & Francis Group.

Wiegand, J. (2014). *The acquisition of receptive L2 grammar knowledge in primary school immersion*. Unpublished Master's Thesis. University of Hildesheim.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language teaching, learning and testing* (pp. 186–210). London: Longman.

Wijnen, F., Kempen, M., & Gillis, S. (2001). Root infinitives in Dutch early child language: an effect of input? *Journal of Child Language*, *28*, 629–660. http://doi.org/10.1017/S0305000901004809

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu: Second Language Teaching and Curriculum Center.

Wright, C., & Tavakoli, P. (2016). New directions and developments in defining, analyzing and measuring L2 speech fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 73–77. http://doi.org/10.1515/iral-2016-9990

Yuan, F., & 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.

6 Appendices

The appendix is freely available at http://dx.doi.org/10.18442/084 upon request.

An increasing demand for English instruction around the globe has long since accompanied large numbers of teachers who speak English as a foreign language. Yet rarely have teachers been taken into account as foreign language speakers, and research on the relationship between teachers' second language performance and the students' acquisition of English is scarce. This doctoral dissertation shows how English teachers at selected German elementary schools perform in English, exemplifies how nine to ten-year-old fourth-graders progress in receptive English grammar and vocabulary, and finally analyzes the relationship between the teachers' English and their students' English development. The results lay out the interrelationships between the complexity, accuracy, and fluency (CAF) dimensions in the teachers' second language production and furthermore demonstrate how features in the linguistic input specifically relate to the students' receptive development of English grammar.

ISBN 978-3-96424-025-5