JYU DISSERTATIONS 682

Ghulam Abbas Khushik

How do the Common European Framework Levels Differ in Terms of Linguistic Features?

Analysing English Language Learners' Written Corpora by Using Natural Language Processing Tools

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF HUMANITIES AND
SOCIAL SCIENCES

JYU DISSERTATIONS 682

Ghulam Abbas Khushik

How do the Common European Framework Levels Differ in Terms of Linguistic Features?

Analysing English Language Learners' Written Corpora by Using Natural Language Processing Tools

Esitetään Jyväskylän yliopiston humanistis-yhteiskuntatieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi Agoran auditoriossa 3
syyskuun 2. päivänä 2023 kello 10.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities and Social Sciences of the University of Jyväskylä,
in building Agora, auditorium 3, on September 2, 2023, at 10 o'clock.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2023

Editors
Ari Huhta
Center for Applied Language Studies, University of Jyväskylä
Päivi Vuorio
Open Science Centre, University of Jyväskylä

Copyright © 2023, by the author and University of Jyväskylä

ISBN 978-951-39-9706-9 (PDF)
URN:ISBN:978-951-39-9706-9
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-9706-9

ABSTRACT

Khushik, Ghulam Abbas
How do the Common European Framework levels differ in terms of linguistic features? Analysing English language learners' written corpora (by) using Natural Language Processing tools
Jyväskylä: University of Jyväskylä, 2023, 174 p.
(JYU Dissertations
ISSN 2489-9003; 682)
ISBN 978-951-39-9706-9 (PDF)

The Common European Framework of Reference (CEFR) for language learning, teaching, and assessment developed by the Council of Europe (CoE, 2001) is a widely used reference source to increase transparency in language education across Europe. It is a helpful document across European countries, particularly for testing and examination providers (Martyniuk, 2005). The CEFR is "one of the most ambitious examples of the gradual formation, shaping, and... implementation of language education policies" (Byrnes, 2007, p. 641). The CEFR provides a basis for elaborating language syllabuses, curriculum guidelines, examinations, and textbooks across Europe and beyond. It has become a focus of interest for two research areas: second language acquisition (SLA) and language testing. SLA research is interested in understanding how second and foreign language proficiency develops, whereas language testing research is interested in developing valid and reliable measures to assess language proficiency.

However, the CEFR has a limited empirical basis in learner languages (Alderson, 2007; Hulstijn, 2007; Wisniewski, 2017). For example, the linguistic features of learners of English as a foreign language (EFL) regarding the CEFR scales are not explored extensively. The CEFR refers to the linguistic features and complexity in learners' language in the descriptors in the scale called 'General Linguistic Range,' for example, but does this in a very general manner. Linguistic complexity in the CEFR includes references to syntactic complexity in terms of complex, simple, and basic structures but, overall, syntactic complexity as described in the CEFR descriptors (i.e., the 'Can-do' statements) remains defined only at a very general level. The focus of the CEFR scales is clearly to describe the ability to communicate and function using the language rather than to describe which syntactic or other linguistic features are typically needed when learners use a particular foreign language. This is understandable since the CEFR is intended to be applicable across different languages. This study paves the way for describing the linguistic basis (syntax, more specifically) of the CEFR levels by investigating syntactic complexity (SC) in the writings of Finnish-speaking EFL learners from Finland and Sindhi-speaking EFL learners from Pakistan across the CEFR levels A1 to B2. The study yielded quantitative information about different SC features that characterise the EFL learners and the texts they wrote (e.g., the mean values for each SC feature at specific CEFR levels and variation in these values). Such information enables us, thus, to characterise numerically syntactic complexity in the English language texts written by the two EFL learner groups and to investigate if specific indices of SC distinguish different CEFR levels from one another and if the two learner groups differ in this respect. These quantitative results are not directly transformable into verbal descriptors of syntax at different CEFR levels for these L1 background EFL learners. However, such results are likely to assist in such endeavours in the future.

Given the multi-dimensional nature of syntactic complexity, this study investigated a wide range of syntactic complexity features in several writing tasks written by learners in Finland and Pakistan. Trained raters rated learners' scripts on the CEFR levels. Two automated applications (L2 Syntactic Complexity Analyzer and Coh-Metrix) were used to analyse the texts for extracting the syntactic complexity features. Statistical analyses revealed differences but similarities in the syntactic complexity indices between different CEFR levels in the EFL learners' writing in the two countries, representing different linguistic and educational contexts. The results shed light on to what extent the communicatively equivalent CEFR levels are also linguistically similar when the target language is the same but when the learners' first language differs. The study also provides information regarding differences in syntactic complexity related to learners' age or educational level. Finally, the findings may be used in systems that evaluate texts written by EFL learners with reference to a variety of linguistic features in order to predict the CEFR level of these texts.

Keywords: Syntactic Complexity, SFL writing, EFL learner Corpus, the CEFR, Automated Applications

TIIVISTELMÄ

Khushik, Ghulam Abbas
Miten Yleiseurooppalaisen viitekehyksen tasot poikkeavat lingvistisiltä piirteiltään? Analyysi englannin kielen oppijoiden kirjallisista korpuksista käyttäen luonnollisen kielen käsittelyn välineitä.
Jyväskylä: Jyväskylän yliopisto, 2023, 174 s.
(JYU Dissertations
ISSN 2489-9003; 682)
ISBN 978-951-39-9706-9 (PDF)

Euroopan neuvosto on kehittänyt Yleiseurooppalaisen viitekehyksen (EVK; engl. *the Common European Framework of Reference, CEFR*) edistääkseen eurooppalaisen kielenopetuksen läpinäkyvyyttä (CoE, 2001). Eurooppalaista viitekehystä käytetään laajasti kielten oppimisessa, opettamisessa ja arvioinnissa, ja se on hyödyllinen erityisesti arviointien laatijoille (Martyniuk, 2005). Viitekehystä pidetään yhtenä esimerkkinä kielikoulutuspolitiikan kunnianhimoisimmista hankkeista (Byrnes, 2007). Se tarjoaa pohjan kielten opetusohjelmien oppimäärien, opetussuunnitelman perusteiden, kokeiden ja oppikirjojen kehittämiselle Euroopassa sekä muualla. Viitekehyksestä tehdään tutkimusta toisen kielen oppimisen ja kielitaidon arvioinnin aloilla. Toisen kielen oppimisen tutkimus pyrkii ymmärtämään, miten toisen ja vieraan kielen kielitaito kehittyy, kun taas kielitaidon arvioinnin tutkimuksen tavoitteena on laatia päteviä ja luotettavia kielitaidon arviointikeinoja.

Eurooppalaisella viitekehyksellä on kuitenkin rajallinen empiirinen perusta oppijankielten suhteen, sillä kullekin EVK-tasolle tyypillisiä lingvistisiä piirteitä esimerksi englantia vieraana kielenä oppivien kielessä on tutkittu varsin vähän (Alderson, 2007; Hulstijn, 2007; Wisniewski, 2017). Vaikka EVK viittaa oppijankielen lingvistisiin piirteisiin ja kompleksisuuteen taitotasokuvauksissaan (esim. *General Linguistic Range*), se kuvaa niitä vain hyvin yleisellä tasolla. EVK-tasot keskittyvätkin kuvaamaan kykyä kommunikoida ja toimia kieltä käyttäen sen sijaan, että ne kuvaisivat sitä, mitä syntaktisia tai muita lingvistisiä piirteitä tyypillisesti tarvitaan oppijoiden käyttäessä tiettyä vierasta kieltä.

Tämä tutkimus keskittyy tarkastelemaan syntaktista kompleksisuutta kirjoituksissa, joiden tekijät ovat EVK-tasoilla A1–B2 olevia englantia vieraana kielenä opiskelevia suomenkielisiä oppijoita Suomesta ja sindhinkielisiä oppijoita Pakistanista. Koulutetut arvioijat arvioivat oppijoiden kirjoittamat tekstit EVK-tasojen mukaisesti, ja kahta luonnollisen kielen käsittelyn (engl. *Natural language processing, NLP*) automatisoitua sovellusta käytettiin analysoimaan tektien syntaktisen kompleksisuuden piirteitä. Tutkimus antaa määrällistä tietoa erilaisista syntaktisen kompleksisuuden piirteistä, jotka määrittävät englantia vieraana kielenä oppivia ja heidän tekstejään. Tilastolliset analyysit osoittivat sekä eroja että yhtäläisyyksiä syntaktisen kompleksisuuden piirteissä eri EVK-tasoille sijoittuvien oppijoiden kirjoituksissa kahdessa tutkimuksessa mukana olleessa maassa. Tulokset selventävät, missä määrin kommunikatiivisesti toisiaan vastaavat EVK-tasot ovat myös kielellisesti samankaltaisia, kun kohdekieli on sama mutta oppijoiden ensikieli eri. Tutkimus antaa lisäksi tietoa syntaktisen kompleksisuuden eroista oppijoiden ikään tai koulutustasoon liittyen. Tuloksia voidaan käyttää myös järjestelmissä, jotka arvioivat englantia vieraana kielenä oppivien kirjoittamien tekstien erilaisia kielellisiä piirteitä ja pyrkivät niiden perusteella ennustamaan tekstien EVK-tasoa. Tämän ohella tulokset voivat tulevaisuudessa auttaa kuvaamaan EVK-tasojen lingvististä perustaa.

Avainsanat: syntaktinen kompleksisuus, kirjoittaminen toisella ja vieraalla kielellä, englanninoppijoiden oppijankielen korpus, EVK, automatisoidut sovellukset

Author's address

Ghulam Abbas Khushik
Center for Applied Language Studies
Agora, Building Ag, 5th floor, Mattilanniemi 2 PO Box 35 FI-40014 University of Jyväskylä
ghabkhus@gmail.com

Supervisors

Professor Ari Huhta
Center for Applied Language Studies
University of Jyväskylä

Professor Riikka Alanen
Department of Teacher Education
University of Jyväskylä

Reviewers

Xiaofei Lu
Professor of Applied Linguistics and Asian Studies,
Department of Applied linguistics
The Pennsylvania State University, USA

Kristopher Kyle
Associate Professor, Language Teaching Studies,
Linguistics, College of Arts and Sciences
University of Oregon, USA

Opponents

Xiaofei Lu

FOREWORD

I would like to express my profound gratitude to God for granting me the opportunity to achieve my research objectives. I am immensely appreciative of the invaluable contributions made by numerous individuals towards the completion of my dissertation. I am deeply humbled and fortunate to have been inspired and motivated by the following individuals: foremost, I extend my heartfelt appreciation to my esteemed supervisors, Professors Ari Huhta and Riikka Alanen, for their exceptional collaboration on this project. I am also indebted to my committee members for their unwavering patience, consistent guidance, prompt feedback, unwavering support, and immense encouragement throughout the dissertation process.

My sincere appreciation also goes to the pre-examiners for their constructive criticism and recommendations, which proved instrumental in enhancing the quality of my dissertation. Their professional input was invaluable in enabling me to present my conclusions more effectively.

Furthermore, I would like to take this opportunity to express my gratitude to the members of the Center for Applied Language Studies (CALS/SOLKI) who were always courteous and helpful to me. Additionally, I am grateful to the Center for Applied Language Studies and the Faculty of Humanities at the University of Jyvaskyla for their scholarship, which enabled me to undertake my PhD studies.

I am also appreciative of the administrations of the different schools, the parents of the students, and my friends who provided support with data collection. Lastly, I extend my heartfelt gratitude to my family and closest friends for their unwavering encouragement and support, particularly my mother and father, who always offered their best wishes and prayers. Additionally, I remain indebted to my family members and children for their support.

Jyväskylä 30.6.2023
GhuLam Abbas Khushik

FIGURES

TABLES

CONTENTS

1 INTRODUCTION

The present dissertation delves into an investigation of the linguistic foundation of the Common European Framework of Reference (CEFR; Council of Europe 2001; 2020), which holds a significant status in shaping second and foreign language (SFL) learning across European nations as well as in various countries in Asia, the Middle East, and Latin America. The primary objective of this study is to present a comprehensive analysis of the CEFR's theoretical framework and its influence on SFL education, with the aim of elucidating its blueprint and implementation in diverse settings. The CEFR outlines the principles of language learning and usage in terms of scales and exhaustive descriptions, which are employed to develop language courses, curricula, textbooks, and evaluation procedures. These serve to set the learning objectives for acquiring a second or foreign language.

The CEFR defines proficiency levels based on its scales. Consequently, the CEFR provides users with a common reference point for defining SFL proficiency levels. Essentially, the CEFR functions as a map that helps users to navigate second language acquisition in a clear and consistent manner. CEFR is also a valuable tool for teachers. Teachers can also benefit from the CEFR. This allows them to align their teaching with the learner's needs and provide appropriate instruction. Moreover, it serves as a benchmark that allows teachers to assess the progress of their students. The CEFR is like a roadmap, allowing teachers to plan out the best route towards their students' goals and measure milestones along the way.

The Common European Framework of Reference (CEFR) utilizes a precise scale to determine the proficiency levels of individuals in second foreign language (SFL) acquisition. This framework serves as a reliable and standardized reference point, facilitating clear and consistent navigation of the language-learning process for both teachers and learners. Teachers can use the framework to tailor their instruction to the specific needs of their students, monitor progress, and chart out the most effective route towards achieving their students' goals while also measuring critical milestones along the way. In summary, the CEFR is an indispensable tool for anyone seeking to enhance their SFL skills.

In the realm of applied linguistics, scholars have recently posited that collaborative research efforts that integrate language learning and language testing could yield more effective techniques for language instruction and assessment (Bachmann and Cohen, 1998; Hulstijn et al., 2010). Through a thorough examination of the similarities and distinctions between these two fields of research, academics can acquire knowledge that can be applied to enhance language instruction and assessment techniques. Such efforts could result in more efficient language teaching and learning outcomes. Additionally, language testing insights can be utilized by scholars in the domain of second language acquisition (SLA) to obtain more dependable data on language learners' capabilities, and to create more efficacious language learning materials.

Researchers in the field of language testing stand to greatly benefit from the insights provided by SLA research data. This information can enhance their test development processes and enable them to create more precise and accurate language assessment tools. By leveraging the knowledge gained from SLA research, these experts can create more valid tests and rating scales that aid individuals in advancing their second and foreign language abilities. It is worth noting, however, that not all language testing researchers possess the necessary expertise to effectively utilize SLA data or may have alternative research objectives that preclude interest in this information.

To address these challenges, a multifaceted approach involving SLA, language testing, and corpus linguistics research is required to provide a comprehensive view of language proficiency. This approach can inform instruction and assessment, facilitating a more precise and nuanced understanding of language acquisition's complexity and leading to more efficient and effective language learning. Additionally, this approach can help identify the connections between language proficiency, learning, and instruction, leading to the creation of better language learning strategies. It is essential for additional SLA researchers to collaborate with language testing researchers to ensure these studies are conducted in a meaningful manner that leverages SLA data and applies to real-world situations.

The field of language education and research has undergone two significant developments since the beginning of the third millennium. The first is the introduction of the Common European Framework of Reference in 2001, which established a standard for language learning across Europe and led to more consistent language assessment and teaching strategies. The second is the emergence of diagnostic testing as a field of study and research in the early 2000s, enabling teachers to create more targeted and effective language learning plans by comprehensively evaluating language learners' strengths and weaknesses.

These developments have sparked increased interest in the linguistic features of learners' performances that represent different levels of proficiency. Consequently, there has been a greater emphasis on the accuracy and complexity of language learners' language use, leading to a refinement of the proficiency levels established by the CEFR. Overall, these developments have significantly

contributed to the advancement of language education and research, and they continue to shape the field today.

The Council of Europe achieved a ground breaking feat in language education, both within Europe and globally, with the introduction of the Common European Framework of Reference for Languages (CEFR) in 2001. This initiative has significantly enhanced the practical and action-focused methods of language learning, teaching, and evaluation. The efficiency of CEFR can be largely attributed to its proficiency scales, which outline the language learning and usage abilities that learners and users can attain at different proficiency levels. These scales facilitate the definition of language learning goals and standards (levels of difficulty) for educational materials, curricula, and evaluation. As a result, language learners, teachers, and users can better comprehend language learning objectives and track progress more accurately. Furthermore, CEFR scales provide an opportunity to measure language proficiency, enabling learners to identify areas requiring improvement to achieve a specific level of proficiency.

While the Common European Framework of Reference for Languages (CEFR) has been widely accepted, it has also faced criticism. One of the main criticisms is that its levels are based on teacher judgments of descriptors compiled from various scales and tests rather than solid research on second language acquisition. However, it is important to note that some argue that these criticisms have limitations and issues, as noted by North (2020). It is imperative to note that the Common European Framework of Reference for Languages (CEFR) does not provide an exhaustive description of the linguistic competencies needed for effective communication in distinct activities, contexts, or proficiency levels. Nevertheless, the inclusion of these features in language materials and assessments can facilitate the development of content that aligns with each level appropriately. The need for further research to identify the specific characteristics of CEFR levels has been recognized, and the University of Jyväskylä has been actively involved in this through various projects, including CEFLING, TOPLING, and DIALUKI, as well as its participation in the SLATE network.

According to Hulstijn, Alderson, & Schoonen (2010, p. 18), an investigation into borderline language features, identification of common features shared across multiple levels, and development of appropriate scales is imperative for determining the typical language features of a specific CEFR level. In addition, the authors query the efficacy of linguistic characteristics discovered through profiling research as tools for identifying learners' proficiency levels and areas requiring additional attention and training. The English Profile project has conducted research on the linguistic characteristics of CEFR levels in the English language based on learners' spoken and written performances in the Cambridge English language examinations (Hawkins & Filipovic, 2012; Green, 2012).

The present dissertation study is a recent addition to the growing body of international research. However, it diverges from previous studies, particularly in its quantitative approach to the investigation of linguistic features. This approach does not readily elicit the qualitative information (such as concrete

linguistic feature profiles or descriptive scales) that Hulstijn et al. (2010) advocate for, as previously cited. The study focuses on syntactic complexity (SC), a linguistic phenomenon that is not extensively delineated in the Common European Framework of Reference (CEFR) scales, by examining the numerical values of the various SC indicators in the writing of EFL learners. Ortega (2003, p. 492) defines syntactic complexity as "the range of forms that surface in language production and the degree of sophistication of such forms." Syntactic complexity is a vital aspect of complexity, and comprehending complexity is crucial for acquiring, developing, and testing foreign languages.

Interest in the linguistic characteristics of language learner performances in general and at different CEFR levels has also increased. This is because of the rise of diagnostic testing of foreign and second language (SFL) skills as an emerging strand of language testing research. Diagnostic testing tries to identify and understand SFL learners' strengths and weaknesses, which often involves assessment and analysis of learner performances in some detail (e.g., Alderson, 2005; Harding, Alderson & Brunfaut, 2015; Lee, 2021). Thus, understanding the linguistic basis of the CEFR levels is of considerable interest to diagnostic language testers, too, which is why they often collaborate with SLA researchers (see the discussion of the SLATE network above, which was a joint venture between SLA and language testing researchers; Bartning et al., 2010). The University of Jyväskylä and its Center for Applied Language Studies have established diagnostic testing as a new field in applied language studies. It coordinated the development of DIALANG, the first major international language test designed to provide users with diagnostic feedback rather than certificates (Alderson, 2005; Alderson & Huhta, 2005). It has also conducted cutting-edge research on diagnosing SFL skills in the project DIALUKI, from which part of the dissertation data comes (see, for example, Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015; www.jyu.fi/dialuki). As mentioned above, one of the objectives of such diagnostic research is to examine the linguistic components of learners' reading, writing, speaking, and listening abilities at different proficiency levels. This is in a second or foreign language. Thus, the current research is linked to studies on diagnostic SFL assessment.

Considerable research has been undertaken to investigate the link between syntactic complexity and the aptitude of SFL learners, leading to the identification of a robust correlation. However, prior studies have primarily focused on a limited number of syntactic complexity measurements, such as sentence and T-unit length, when analyzing SFL development and proficiency in learners (Ortega, 2003; Norris & Ortega, 2009; Wolfe-Quintero et al., 1998). Furthermore, researchers have often inaccurately assessed learners' proficiency levels by relying on proxies like grade level or age (Carlsen, 2012; Norris et al., 2009; Ortega, 2003).

This dissertation delves into the fundamental linguistic underpinnings of CEFR by investigating syntactic complexity features in the written works of EFL learners from Pakistan and Finland. The study employs close to 30 SC indices

that are extracted using two natural language processing (NLP) automated applications in transcripts that are rated based on CEFR levels.

1.1 Aims and goals

This dissertation delves into the disparities in syntactic complexity among learners categorized under different CEFR levels, such as A1, A2, B1, and B2. The study employs around 30 syntactic indices to scrutinize EFL learners' syntax and comprehend the linguistic groundwork of the Common European Framework levels in English as a foreign language. The research focuses on examining the role of the first language (L1) concerning CEFR levels and the proficiency level of EFL students. The results indicate that the syntactic complexity of learners increases as their proficiency level increases at CEFR levels. Moreover, the study establishes that the syntactic complexity of learners at CEFR levels A1, A2, and B1 is significantly influenced by their L1 and age. Overall, this research presents a comprehensive understanding of the linguistic basis of CEFR levels and the application of syntactic indices to evaluate the complexity of EFL learners' syntax.

The primary focus of the first study was directed towards the learners hailing from two distinct contexts, namely Pakistan and Finland. The linguistic characteristics of the texts were subjected to meticulous examination in order to establish the correlation between syntactic complexity and Common European Framework of Reference (CEFR) levels. The study delved deep into the inquiry of whether the syntactic complexity of learners varied across the A1, A2, and B1 CEFR proficiency levels, and further investigated the role of learners' first language (L1) in the relationship between syntactic complexity and CEF levels. The main objective of the study was to determine whether syntactic complexity (SC) differed across two L1 groups of EFL learners, Finnish and Sindhi, whose CEFR writing levels were similar. The Common European Framework, a set of guidelines utilized to describe language proficiency levels, provided the framework for the study. The term L1 refers to a person's first language, the one they acquired at birth. Examining the relationship between L1 and syntactic complexity is akin to exploring the relationship between height and weight, as a person's L1 may influence their syntax just as height may influence weight. The study examined syntactic complexity as a multidimensional construct using 30 distinct indices. The current research utilized data from eighth- and 12th-grade gymnasium learners in Finland. Additionally, data from eighth- to 12th-grade learners from Pakistan was gathered, scrutinized, and analyzed. The study employed a data set consisting of essays on an argumentative topic rated A1, A2, or B1.

The second study explored Finnish EFL learners' syntactic complexity features. It included 8th graders (aged 14) and 12th-grade upper-secondary school students (aged 17 or 18) who completed three English writing tasks from an international English language examination and a previous research project (see Tasks in Section 6.2.2 for details). Learners expressed their views narratively,

argumentatively, and descriptively. This study explored what SC features correlate with learners' overall writing ability based on three writing tasks. Also, it investigated whether the correlation with SC features was different for 8th and 12th graders across A1, A2, and B1 levels. The second main aim was to examine which SC indices distinguish between the CEFR levels covered in the study. In this regard, Study 2 was like Study 1, except that it focused only on the L1 group. However, the fact that Study 2 included two age groups (or groups whose educational levels differed) also allowed some limited investigation of possible age-related differences in the SC indices that separated the CEFR levels. To ensure that the results of the study were not simply due to the educational level of the participants, Study 2 was designed to compare the SC indices in two separate age groups; this allowed researchers to better understand which SC indices could be used to distinguish between different CEFR levels regardless of age.

The third study explored the syntactic complexity features of two different age groups but with the same CEFR proficiency level. This study was conducted with data sets from Pakistan and Finland. In previous studies, none focused on analysing learners with an identical CEFR level and attempted to answer whether learners, for example, at A1 proficiency from the eighth vs 12th grade, have similar syntactic complexity features (or, more precisely, the same values for these SC features). The purpose of this study was to explore whether learners of the same proficiency level, but from different age groups, have similar syntactic complexity features. The data sets used in this study come from two different countries, Finland and Pakistan, which provides a unique opportunity to compare those features across different contexts.

Overall, the findings of the current dissertation are valuable in understanding how SC features (their numerical values) relate to proficiency levels, particularly from the point of view of how SC changes as learners move from one level to the next. Multivariate analyses showed significant variations across CEFR levels. As did subordination indices, production unit indices varied substantially between CEFR levels in both L1 and age groups. Similarly, significant differences were found for working memory load, referring expressions, syntactic variety and simplicity, and phrasal density across CEFR levels. These differences were found between Pakistanis and Finns. This suggests that the CEFR levels are a useful tool for measuring the differences in language proficiency across different countries and age groups. Moreover, the results of the multivariate analyses indicate that the CEFR levels have a significant effect on the complexity of language production.

This research has paved the way for further research in the future to explore *which linguistic features are different across the CEFR levels and which features are typical at each level*. Thus, this research provides a basis for the CEFR levels to be better understood and for further research to identify the specific language features that are associated with each CEFR level. This could help educators to better understand and teach language at each level, as well as provide language

learners with an understanding of the language features they need to acquire in order to reach a certain CEFR level.

A strong correlation exists between proficiency levels and Second Language (L2) features (see Ortega 2003 & Wolfe-Quintero et al. 1999). Since the CEFR provides a framework for assessing learners' proficiency in foreign languages, and an assumption is that samples rated at the same CEFR level may have similar syntactic complexity. However, there may be significant differences in L2 features. A number of factors may contribute to these differences, including the learners' native language (L1) or their age. Children, for example, tend to acquire L2 features quickly due to their natural language learning abilities, and may thus demonstrate more features typical of a higher CEFR level than their actual competency. As a result, older learners may be slower to acquire L2 features, resulting in a lower CEFR level than what their L2 features would suggest. The differences in L2 features that result from L1 or age can also be observed in the classroom, as some students progress at a faster rate than others. It is important for teachers to be aware of these differences and adjust their teaching methods accordingly. In addition, teachers are advised to take into account the learners' L1 and age when assessing their proficiency as this can give a more accurate indication of the learners' actual level of proficiency.

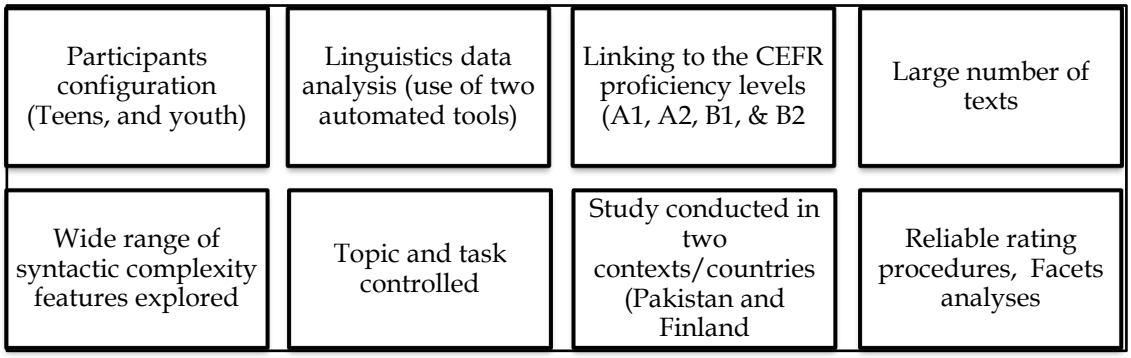
1.2 Important features of the study

This section summarises the key features of the study, some of which are rare in previous research on syntactic complexity.

Age of the informants: One of the gaps in SC research has been the age of the learners. The current study investigated syntactic complexity in younger learners of English as a foreign language (EFL) than in most previous studies. The participants were teenagers and young adults from Finland and Pakistan in the 8th, 9th, 10th, 11th, and 12th grades.

Automated analysis of texts: Coh-Metrix (Graesser et al., 2004) and L2 Syntactic Complexity Analyzer (Lu, 2010) are automated applications that analyze written scripts to produce multi-dimensional complex features.

Figure 1.1 Key features of the current dissertation



Reliable placement on proficiency levels: In our data, the learners' proficiency levels have been determined based on Common European Framework of Reference (CEFR) levels (A1, A2, B1, and B2) through multiple ratings by trained raters. The researcher analysed the reliability of ratings and raters in the multifaceted Rasch program Facets, thus applying language testing expertise to ensure that the data gathering and assignment to the proficiency levels were as high quality as possible.

The CEFR levels: The study reports on the syntactic complexity feature distinguishing learners' CEFR proficiency levels. The resulting numerical characterisation of the level (and their differences) can help researchers define the linguistic profiles of the A1-B2 levels for Finnish and Pakistani (Sindhi) EFL learners in terms of syntax. Of particular importance here is the fact that the present study also covered the A1 level, which has not often been investigated in CEFR-related research on SC.

Topic and task were controlled: Attention was paid to the comparability of the data-gathering procedures in both countries and the choice of the task(s) to be included in the analyses (this is another feature where the procedures used in language testing research were applicable). When the two L1 groups were compared, the comparison was based on precisely the same task and topic.

A large number of SC indices: The multidimensional analysis of syntactic complexity in EFL learners' writings using roughly thirty syntactic features reveals a complete picture of syntactic complexity at each competence level and syntax progression (expressed as numerical indices) across the CEFR proficiency levels. Probably the largest-ever investigation of the SC measures on learner writing related to specific topics, grades, and competence levels is one of the strengths of this study compared with past studies. As a result, this research presents an extraordinarily detailed (numerical) picture of syntactic complexity in EFL writing. It also reveals how learners' first language and age affect syntactic complexity.

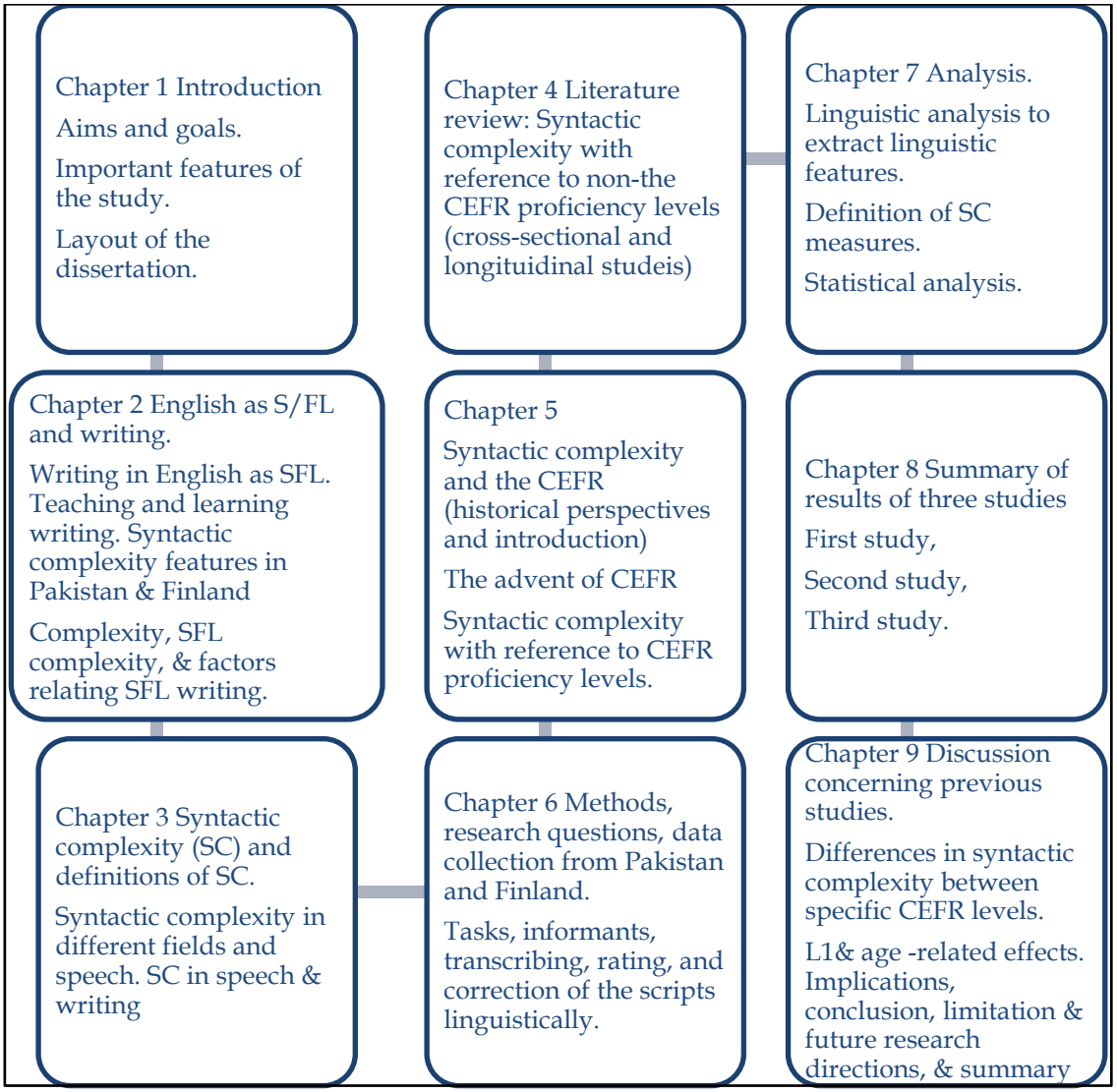
In our current study, we utilized prior research on second/foreign language syntactic complexity (SC), specifically studies regarding CEFR levels, to identify key features that are crucial for beginning and intermediate-level learners. Drawing from evaluations by renowned scholars like Ortega (2003) and Wolfe-

Quintero et al. (1999), we carefully selected SC indices that have been empirically proven to be reliable indicators of syntactic complexity in second/foreign language writing. Our primary aim was to choose indices that have already been validated in past research, thereby ensuring the accuracy and relevance of our findings.

1.3 Outline of the dissertation

The dissertation's second chapter focuses on English as a second/foreign language (SFL) and SFL writing. The third chapter focuses on syntactic complexity, including its definitions and investigations in numerous areas and disciplines. The fourth chapter presents syntactic complexity regarding proficiency levels that must be aligned (at least not directly) with the CEFR level. The fifth chapter presents a literature review on syntactic complexity with reference to the CEFR levels. Chapter 6 elaborates on the methodology used in this dissertation. The various analyses are discussed in Chapter 7. Chapter 8 presents a summary of the three sub-studies comprising the dissertation. Discussion, consequences, limitations, future directions, conclusion, and a summary comprise Chapter 9.

Figure 1.2. Graphical outline of the dissertation



2 ENGLISH AS A SECOND / FOREIGN LANGUAGE (SFL) AND SFL WRITING

The English language and culture have attained worldwide prominence. English as a second or foreign language is becoming increasingly essential worldwide. The physical presence of English colonists in Australia, America, Asia, Africa, and the Middle East strengthened and reinforced the significance of the English language, as did the translation of foreign literature into English. In many contexts, English individuals invented the alphabet and published foreign grammar books (e.g., Trump, 1866) and alphabets, such as the Sindhi alphabet, which the British government produced in 1852. Thus, English as a second or foreign language has become increasingly essential worldwide.

Non-native speakers of the English language are taught English as a second or foreign language at all stages of education, from elementary school through college and university, even though it can also be challenging to teach English in various situations across the globe. Generally, organizing classes for non-native English speakers posed a significant issue for natives in the past. In order to solve this issue, early investigations concentrated mainly on SFL education programmes (e.g., Agard and Harold, 1948).

English as a second or foreign language instruction and assessment are interrelated fields. It is essential to comprehend the context in which SFL English is taught, learned, and evaluated. English was initially taught to English speakers as a foreign language using the Grammar Translation Method, emphasizing English grammar (Low & Lond, 1897; Mulligan, 1853 & 1868). In 1878, Berlitz schools, on the other hand, implemented a direct approach to language training.

In contrast, the "Reform Movement" took place somewhat later. The reform attempt began with the inclusion of phonetics, public speaking, and essay writing, but it failed to gain favour (Jesperson, 1905; Wyld, 1907). Regarding the education of non-native English speakers, English grammar remained an important subject in school debates, mainly around the Educational and Literacy Bulletins (Alderman 1927, & Talbot 1916). Later, Palmer (1921) pioneered oral strategies for English language learning, while Bloomfield pioneered structural-situational approaches to English language instruction (1926).

Teaching English as a second or foreign language was of primary importance, followed by procedures and tests for evaluating English as a second or foreign language (Agard & Harold, 1948; Bloomfield, 1926 and 1933). This brief history of EFL instruction and evaluation illustrates the approaches that existed from the beginning and may help explain why there are so many approaches in contemporary EFL education.

2.1 English as a second and foreign language in Pakistan and Finland

2.1.1 Brief introduction into writing in English as SFL

This section will discuss the importance of writing skills in European countries. After the 1960s decline of the "Grammar Translation Method" in foreign language instruction, European countries neglected writing skills for an extended period. Nonetheless, writing gained popularity in the classroom and in research during the 1980s and 1990s, and there is now a flourishing dialogue addressing theoretical and practical issues and problems in L1 and SFL writing. Second language writing became a focus and a multidisciplinary field of study encompassing a variety of academic and educational situations (e.g., Paul & Pew, 2002, p. 261).

The initial focus of the discussions on writing was on theoretical and practical challenges and methodological developments. The writing researchers initially investigated the theoretical, practical, and practical obstacles and problems (et al., 1990) and the methodological changes in teacher training for writing instruction (Wolfgang, 1995). Due to methodological developments, a previous study on writing instruction produced two contradicting conclusions. Some rejected writing instruction, while others supported it.

2.1.2 Writing in English language in Pakistan

This section discusses the English language in Pakistan before going into the features of written Pakistani English. English is spoken in around 46 dialects on the world's continents. The grammatical properties of these variants may vary. In the British Isles, American, Caribbean, Pacific, Australian, and African or Asian varieties of English, for example, there are varying proportions of the noun phrase, verb phrase, and negation (see Kortmann et al., 2004, p. 1194). Some linguistic characteristics of Asian English dialects (such as Pakistani English) cannot be found elsewhere. In Pakistani English, for instance, numerous comparatives, superlatives, and progressives are standard (Mahboob, 2004). The Pakistani English language diverged from its beginnings in terms of syntactic structures and word selection at various places in sentences, distinguishing itself from native English varieties. Numerous varieties of Pakistani English have developed, and some linguistic challenges faced by people learning Pakistani

English have been linked to the enormous range of English spoken there (Shahid Muhammad, 2012). Rahman (2015) focused on Pakistanis' syntactic use of English, which demonstrated that the L1 of the Pakistanis might affect their English, for example, regarding the relative proportions of the parts of speech they use in their English.

Other factors, including differences in curricula and textbooks, contribute to the substantial variety in Pakistani English and the ongoing changes. The curricula and textbooks used by English language learners in Pakistan differ by school and study level. According to Nayyar et al. (2005, p. vi), "inconsistent teaching standards and bad English" can hinder the development of English as a foreign language among students. Nayyar et al. (2005) analysed curricula and textbooks produced for English by schools that used them to teach English to students in grades one through twelve. They identified significant variations in the quality of language across all books. There were stylistic flaws, apparent inaccuracies, and pedagogically unsuitable content. In addition, the researchers found "outdated and incoherent instructional approaches that hamper the development of students' attention and insight" (Nayyar et al., 2005, p. vi). They suggested changing these inadequate curricula and textbooks that had been prepared by such uninformed individuals (Nayyar et al., 2005, p. 2).

English writing in the Pakistani context possesses unique linguistic qualities. The syntactic complexity features analysed in this dissertation have undergone several syntactic alterations when the language is used in various cultural contexts and by speakers with a range of native tongues (Anwar, 2009, p. 3). According to Mahmood et al. (2011, p. 50) and Mahmood et al. (2015, p. 412), local linguistic patterns in diverse varieties of English can result in grammatical changes and emerge as modified structures in the forming structures of a new variety of English. To highlight this shift in language, Gardezi (2009) found that Pakistani student essays were much longer and had around 20 words on average, more conjunctive ties per sentence, and ten times more "but" and "and." In most instances, Pakistani pupils must be instructed on how to conclude or begin a sentence and are instead instructed on the correct utilisation of various elements of speech (Gardezi, 2009). In this regard, two studies are significant. Baumgardner (1993) researched newspaper articles, whereas Gardezi (2009) analysed college student essays. Both agreed that Pakistanis write differently from native British speakers of English; this may be attributable to the educational methods and coaching provided in schools. Pakistanis and Britons differed syntactically in the number of words used in essays, clausal modifications, deviation from completion norms, and use of tenses (Mahboob, 2004). In the same vein, Baumgardner (1993b, p. 259) observed that Pakistanis employ fewer infinitives and more gerunds.

According to Shahid (2012), Pakistani pupils often need to gain advanced vocabulary and grammar knowledge and the ability to construct complex sentences and paragraphs. In addition, Pakistani students often need to employ proper punctuation, phrase, and paragraph structure. In addition, writing is rarely emphasised explicitly in the classroom, leading EFL students to imitate

their original language's patterns, structures, and vocabulary selections when translating to English. When translating from their L1s to English, learners usually demonstrate L1 interference, L1 grammar rules, and unusual spellings. The pupils initially think in their original tongue and then translate their thoughts into English, replete with syntax (Muhammad Shahid, 2012).

Studies have shown that nominal and noun phrases are more common in Pakistani than in British and American corpora (Mahmood, 2009, p. 98; Talaat, 2002). Baumgardner (1993) and Mahmood (2009, p. 69) found that Pakistani English speakers replace "to-infinitive verbs" with "that-clause complements" more frequently than British and American corpora. Mahmood (2009) noted that the British and American English corpora have fewer subjunctives than the Pakistani English corpus. Pakistanis may utilise alternative prepositions and the conjunction "and" compared to their British counterparts (Mahboob, 2004, p.1052). Pakistani English (12.64 %) employs more prepositions than British English (11.24%), according to Asghar et al. (2014, p.218).

Numerous syntactic changes occur when a language is used in various cultural contexts. These changes are visible in Pakistani English due to code-switching from the speaker's or writer's native tongue to the English language (Anwar, 2009, pp. 2–3). Additionally, due to norms and standards, local language patterns can change grammatical and evolve into the forming structures of a new variety of English (Mukherjee et al., 2015, p. 412) and Mahboob (2004, p. 1052).

As an independent variable, Pakistani English has also borrowed several terms from the country's indigenous languages (Mahmood, 2009, p. 93). This has some advantages, such as Asghar et al.'s (2014) finding that Pakistani English uses more prepositions than British English, as noted above.

2.1.3 Particular syntactic complexity features of the learners from Pakistan

To conclude the description of the English language in Pakistan, it is essential to discuss the differences between the schools in Pakistan in more detail since such differences explain why the EFL learners from Pakistan in this study are likely to be much more heterogeneous than the EFL learners from Finland. Due to the variety of curricula offered in Pakistani institutions, the learners in the current study are diverse. The unfortunate ones are coerced into following the curriculum the Ministry of Education provides in each province. In comparison, lucky students from prosperous families get their education at O and A levels via English-medium institutions that follow the Cambridge or Oxford curricula (Nayyar et al., 2005, p. 139). Along with the curriculum, teaching techniques at schools vary according to their affordability, size, and location in urban or rural areas. The costliest schools in significant cities follow a curriculum based on books issued by Oxford or Cambridge University Press and teach in English. Another tendency is the development of regional curricula by regional governments, which teach English and other topics in their public schools using a regional or national language. The third kind of school developed its curriculum by combining several books that national and foreign writers had authored. As a result, the language of instruction differs across the institutions.

Additionally, certain schools are administered by the federal government and use the federal curriculum to teach the language. The most complex network is that of private schools, which follow a variety of different curricula.

Nayyar et al. (2005) examined the curricula for teaching English as a foreign language to learners in grades one through twelve. Lesson levels varied across all books in each curriculum. They discovered linguistic differences in the lessons for each grade level across all the books used in various schools. Additionally, English is taught independently of the learner's skills, age, or background. Thus, language development is likely influenced by the learner's age, first language, and socioeconomic circumstances.

2.1.4 English language in Finland

English plays a significant role in Nordic countries. In particular, English is rather popular as a medium of instruction in these countries (see Wächter and Maiworm, 2008a; Weisser, 2016; OECD, 2015, Box c4.1; OECD, 2018). Numerous educational institutions, particularly in higher education, offer instruction in the English language (MOEC, 2018). These institutions have increased the number of English-language higher education programs in Finland and across Europe (Wächter 2008b).

Information gathered in the past few decades indicates that English is studied widely in and outside educational institutions. It is also used quite frequently for various purposes. Regarding the general population's English skills, Kaplan and Baldauf (2005, p. 191) noted that 66% of Finns told an adult education survey in 1995 that they could speak English. The use of the English language has spread in various domains (the media, economics, science, and education); companies have adopted the language increasingly, and study programs in the English language have increased in the Finnish educational system (Kaplan & Baldauf, 2005, p. 216).

When we look at English in the educational system in Finland, we can see a gradual increase over the decades, resulting in English being the most popular foreign language in schools. Unlike in some other countries, English is not a compulsory subject in Finnish schools, but practically everybody starts it at some point, usually as the first foreign language (as the so-called A1 language, which is a compulsory subject; the languages to choose from differ between schools, however).

English as a foreign language has a long history in Finland. English instruction began as early as 1857 in a few private schools (Krook, 1893, p. 110), but it was not until 1886 that it was introduced into some of Finland's public schools. According to Takala (1980, p. 48), only 10,000 people in the 1950s studied a foreign language in Finland. The country is relatively small, and most people recognise the importance of knowing other languages, mainly English nowadays, and are motivated to study them (Ringbom, 2012, p. 495). Initially, German was the primary focus of foreign language education in Finland (before World War II), but the emphasis changed to English after the war. English has grown in popularity in Finland since then. However, not everyone applauded English's

dominance in Finnish educational institutions (see Nikula et al., 2010; Ylönen & Kivelä, 2011), as it has meant a decline in studying other languages. Nonetheless, English has risen to the top of the list of the most frequently spoken foreign languages in Finland, earning it the moniker of Lingua Franca, or a language shared by a large number of Finns, at least in specific professional, educational and free-time contexts (Erickson et al., 2015, p. 204; Hynninen, 2016, p. 71; Jenkins et al., 2018, p. 325; Palviainen, Kalaja & Mäntylä., 2012, p. 2; Ringbom, 2012, p. 495). Similarly, CLIL (content and language integrated learning) has gained popularity, with English being the most frequently used FL in which material is taught across many contexts (primarily high schools).

In general, English usage has increased considerably in Finnish educational institutions. According to a survey published in the Finnish publication "Statistics Finland," 74% of students in grades 1-6 studied English. In contrast, nearly all students in grades seven to nine studied English in comprehensive schools (for more information, see http://www.stat.fi/til/ava/2018/02/ava 2018 02 2019-05-23 en.pdf). According to the survey, English was the most frequently picked language for comprehensive school pupils to master in 2018. Most recently, most pupils chose English as their A1 language (their first foreign language), and nearly 90% of primary school children chose English as their A1 language, which nowadays begins in grade one (Karvi, 2019). English is also gaining popularity at the university level and in practical vocational educational training (VET). Most students enrolled in Finnish VET programs study English in addition to their second national language, Finnish or Swedish (OPH 2019, p. 36, Reports and surveys 2019: 7, Key figures on vocational education and training in Finland). The use of English in Finnish higher education institutions has expanded considerably. English is used in 46% of non-native language degree programs (Garam, 2009, p. 14).

Following WWII, European countries established student exchange programs to boost the English language's development. Numerous educational institutions are Erasmus participants, enabling them to send and receive students from and to European nations. These exchange student (Erasmus) programs require participants to acquire a variety of skills and to improve their ability in the English language as a foreign language (OPH 2017). As a result, the number of educational institutions offering diverse English-language programs continues to expand in Finland and throughout Europe.

Additionally, like in many other countries, Finnish language education has been substantially influenced by the CEFR. Finland was among the first countries to use the CEFR (2001) in curriculum design, language teaching, and language assessment. The Finnish national curricula for comprehensive and upper-secondary schools (OPH 2003 & 2004) specified language learning, teaching, and evaluation targets using a modified version of the CEFR scales. Additionally, in Finland, other language tests, for example, the National Certificates, are based on the CEFR's content and scales (the NC targets adult learners seeking certification of their language skills; see Ahola & Leblay, (2014)). In Finland, the six CEFR levels are divided into eleven sub-levels. The A1 beginner level has three

sublevels (A1.1, A1.2, and A1.3). In contrast, the A2 beginner level has two (A2.1, A2.2), the B1 intermediate level has two (B1.1, B1.2), the B2 intermediate level has two (B2.1, B2.2), and there is one advanced C level called C1.1.

The use of the CEFR in the national curricula for the schools continues. The newest national core curricula for comprehensive and general upper secondary education (OPH 2014, 2015) continue defining targets for learning, including foreign languages, including English, using proficiency levels based on the CEFR levels. The national curricula define which proficiency level the language learner should have at the end of essential stages of schooling to be given a mark of 8 (on a 4 to 10 scale) that corresponds 'good' level of command of the foreign language. These stages are the end of grade 6 (end of primary school), grade 9 (end of comprehensive education), and the end of general upper secondary education (gymnasium). The curriculum defines several dimensions that the teacher should take into account in marking the pupils or students, such as "skills in participating in an interaction," "skills in interpreting texts," and "skills in producing texts." The last one, thus, covers both written and spoken texts. The target level for mark 8 (good command) is the same across all dimensions of proficiency in English in the current national curriculum for grade 9 (where the target is B1.1) and upper secondary schools (where the target is B2.1). However, for grade 6, the target level for mark 8 is slightly lower for writing (i.e., level A1.3 or high A1) than for the other skills (A2.1 or low A2). These target proficiency levels for writing in English as a first foreign language in the Finnish educational system are the following:

- grade 6: level A2.1 (low A2; see OPH 2014, p. 223)
- grade 9: level B1.1 (low B1; see OPH 2014, p. 352)
- end of upper secondary (about grade 12): level B2.1 (low B2; see OPH 2019, p. 177)

Since the national curricula define what kind of proficiency characterises "good" (mark 8) command of English as the first foreign language, the Finnish National Agency for Education (*Opetushallitus* in Finnish, abbreviated as OPH) has published more detailed descriptions for levels of achievement for the end of compulsory education (i.e., end of comprehensive school) in grade 9. These descriptions (OPH 2020) define additional points in the 4 to 10 scale besides the existing definition for mark 8. Definitions are now available for the minimum pass mark (mark 5) and for marks 7, 8, and 9. Hence, for English as the first foreign language, the following proficiency levels define how well the student should be able to write (or read and speak) to be awarded a particular mark on their (comprehensive) school leaving certificate at the end of grade 9 (at the age 15/16):

- mark 5 is awarded if the student has level A1.3 (high A1)
- mark 7 is awarded if the student has level A2.2 (high A2)

- mark 8 is awarded if the student has level B1.1 (low B1)
- mark 9 is awarded if the student has a level that is between B1.1 and B1.2 (OPH 2020, p. 126)

The Finnish educational authorities regularly study student achievement, typically focusing on the end of comprehensive education (grade 9), but foreign language results should be investigated more regularly. The latest large-scale studies with statistically representative samples of learners of English (as the first foreign language) took place in 2013 (for the 9th graders) and 2018 (for the 7th graders).

The 7th graders' study in 2018 found the following results for writing in English (Härmälä et al., 2019, p. 54):

Proficiency level in writing	Percentage of pupils who achieved this level
A1.1 or below	13.1
A1.2	16.2
A1.3	18.3
A2.1	19.9
A2.2	17.7
B1.1 or above	14.8

The 9th graders' study in 2013 found the following results for writing in English (Härmälä et al. 2014, p. 65):

Proficiency level in writing	Percentage of pupils who achieved this level
A1.1 or below	4.5
A1.2	4.6
A1.3	6.7
A2.1	11.9
A2.2	14.6
B1.1	18.3
B1.2	20.4
B2.1 or above	19.0

In the 9th graders' study in 2013, 57.7 % of the students achieved or exceeded the B1 level in writing. In the newer study in 2018, at the beginning of grade 7 (which means that in practice, it tested what the students had learned in primary school), 52.4% achieved at least level A2 in writing. These results are pretty good, but many students cannot write very much in English.

2.1.5 Writing in English (SFL) in the Finnish context:

English language instruction began in Finland in 1857 in the few schools that offered it; textbooks played a critical role in instruction (Krook, 1893, pp. 114–116). Most schools still taught "traditional grammar" (Krook, 1893, p. 116). English language teachers preferred Finnish copies of the grammar texts in private schools over state schools. The teachers emphasised practical grammar in English language instruction (see Krook, 1893, p. 117). Most schools in the Nordic countries concentrated on grammar rules; little attention was paid to the direct approach to teaching English (Dunlop 1953, quoted in Evans 1976). However, teachers explained grammar principles in various ways during their sessions.

Grammar-reading and writing-focused instruction is a more traditional type of instruction than the communicative approach. Reading and writing are still necessary for schools (Korhonen 2010, p. 1, 10). Korhonen (2010, p. 1, 13–15) discovered that students practice their writing skills individually, in pairs, or groups during sessions.

However, the Finnish school system places a greater emphasis on other language abilities than on writing instruction. According to the OECD (2018, p. 340), in Finland, time is allocated to writing instruction for L1 at a rate of roughly 7% in primary school and 4% or less in lower secondary school (p. 3340–3341). English as a foreign language receives approximately 1% of instructional time in primary education and approximately 5% in lower secondary schools. There appears to be no research on how much time is spent practising writing in English classes. However, writing in English is likely practised more in the upper secondary school (Gymnasium) because the Matriculation Examination, the school-leaving examination for that level of education, has written as an essential component of the English (and other foreign/second language examinations). A third of the points in these final tests are based on the student's performance on the writing task. There is no such national examination at the end of the primary, comprehensive education in grade 9; the teachers decide on the final marks, and the weight they place on writing is likely to vary since the national curriculum does not dictate how much emphasis should be given to the various criteria specified in the curriculum (ability to produce spoken and written texts is just one of the several criteria).

2.2 Complexity

The dissertation analysed linguistic scripts produced in English as a foreign language (EFL) to examine and monitor the complexity sub-branch (i.e., syntactic complexity). Nevertheless, it is necessary to comprehend complexity before comprehending its syntactic complexity. Thus, complexity and syntactic complexity are the key theoretical concepts used in this dissertation.

In 1996–1998, Skehan introduced the term "complexity, precision, and fluency" to SLA research. Complexity is a facet of the trinity of complexity,

accuracy, and fluency (CAF), which is a widespread research issue in second language learning and practical linguistics (Housen et al., 2012). Exploring the numerous definitions of complexity and their similarities and differences is necessary. The term "complexity" refers to linguistic and cognitive complexity in the SLA literature (Housen et al., 2012, p. 4; see also DeKeyser 1998; Housen et al., 2012; Housen & Kuiken, 2009; Housen, Van Daele, & Pierrard 2005; Williams & Evans 1998.) Structures that are complicated, difficult to process, and intricate are evidence of their complexity. Complexity, linguistic complexity, and cognitive complexity are often used interchangeably.

The definitions of complexity include:

- "Complexity is commonly characterized as the ability to use a wide and varied range of *sophisticated structures* and vocabulary in the SFL" cf. Ellis 2003, 2008; Ellis & Barkhuizen 2005; Lennon 1990; Skehan 1998; Wolfe-Quintero, Inagaki & Kim 1998). Housen et al. compiled this definition, (2012 p. 2).
- "Complexity refers to the extent to which the learners use syntactic embedding and subordinate clauses, relative to the total number of clauses produced" Housen et al., (2012 p. 4).
- "[complexity is the] use of more *challenging and difficult language* … Complexity is the extent to which learners produce elaborated language" (R. Ellis & Barkhuizen 2005: p. 139).
- "Grammatical and lexical complexity means that a wide variety of both basic and *sophisticated structures* and words are available to the learner" (Wolfe-Quintero, Inagaki, & Kim 1998: p. 69, 101).
- "Complexity refers to … the complexity of the underlying interlanguage system developed" (Skehan 2003: p. 8).
- "Language complexity is a multifaceted, multidimensional and multi-layered construct, a fact that is still insufficiently acknowledged in SFL research" Bulté & Housen 2012 p. 41
- "a matter of the number and variety of an item's constituent elements and of the elaborateness of their interrelational structure" Rescher (1998: p. 1)
- 'A definition of grammatical complexity can be based on the usual understanding of a complex system as one consisting of many different elements each with a number of degrees of freedom' (Nichols, 2009: p. 111)
- 'Complexity should … be defined, to put it in the most general terms, as the number of parts in a system or the length of its description' (Miestamo, 2008: p. 26)
- 'The number of discrete components that a language feature or a language system consists of, and the number of connections between the different components' (Bulté and Housen, 2012: p. 24).

Complexity is partly linked to the learners' explicit, declarative, and implicit procedural interlanguage knowledge. Language development is non-linear, so some characteristics may develop while others do not. Thus, analyzing complexity can help us comprehend the dynamics of language development (Housen et al., 2012; Larsen-Freeman, 2006; Norris & Ortega, 2009; Spoelman & Verspoor, 2010). That is why there are calls for fine-grained analyses of more specific, different complexity measures to complement the global measures (Norris & Ortega, 2009; Robinson & Ellis, 2008). Norris & Ortega (2009) supported the multidimensional approach to studying complexity with previously used elite measures and new fine-grained measures. These measures are essential to capture the complexity at all structural levels, such as sentence, clause, and phrase levels, so that we can look at syntactic complexity multidimensionally.

2.2.1 SFL complexity

The complexity of a second or foreign language is a significant component of developing or acquiring a second or foreign language. It is of considerable importance for the learning, teaching, testing, and research of SFL proficiency. Most studies have referred to SFL complexity as L2 complexity; it is referred to as SFL complexity in this dissertation. There has been substantial research and exploration into the complexity of SFL. Bulté and Housen (2012), Crossley and McNamara (2014), as well as Norris and Ortega (2012) and Ryshina-Pankova (2015), are among the authors who have investigated SFL complexity in writing. In particular, syntactic complexity, which is a component of structural complexity in SFL complexity, has emerged as a critical topic of study in SFL development and acquisition and testing research concerning writing proficiency (Crossley et al., 2014 & 2015; Guo et al., 2013; Lu, 2011 & 2015).

Figure 2.1 A taxonomy of complexity constructs (Bulté and Housen 2012 p. 23)

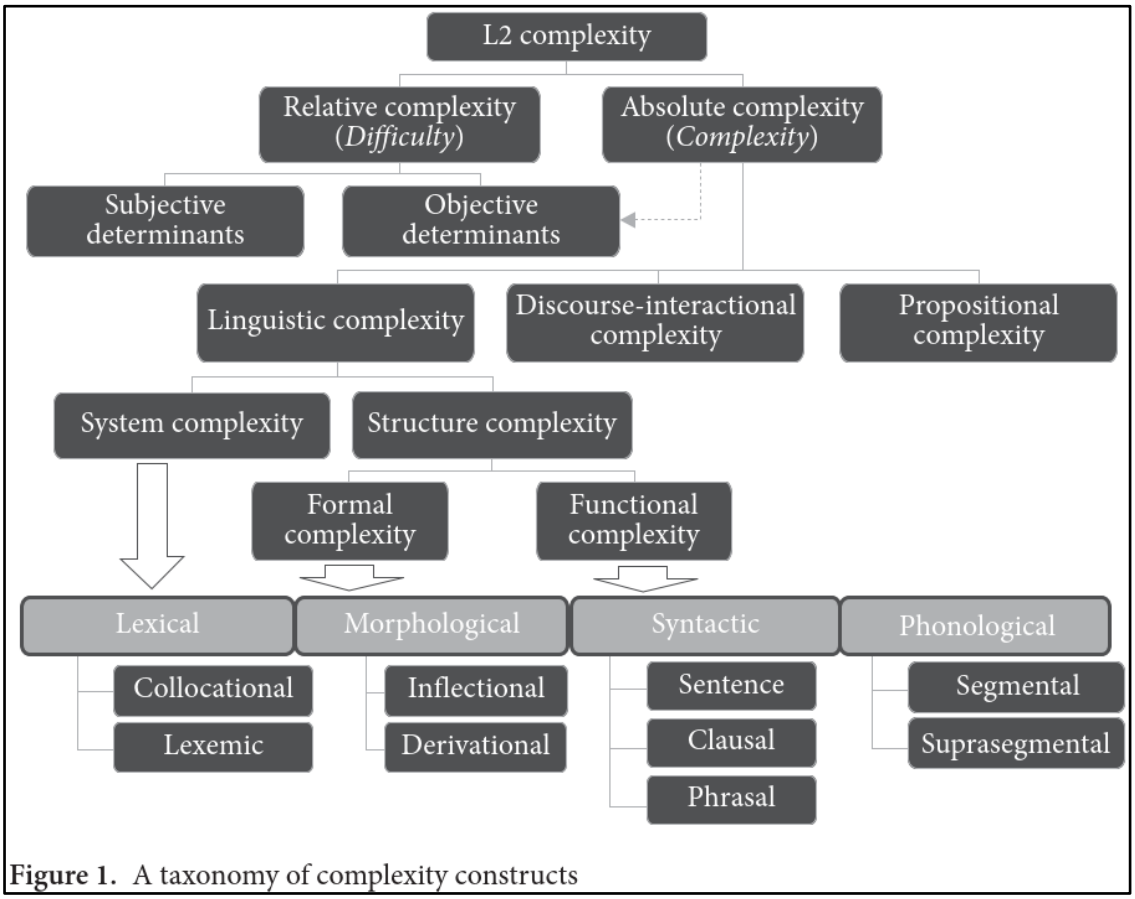
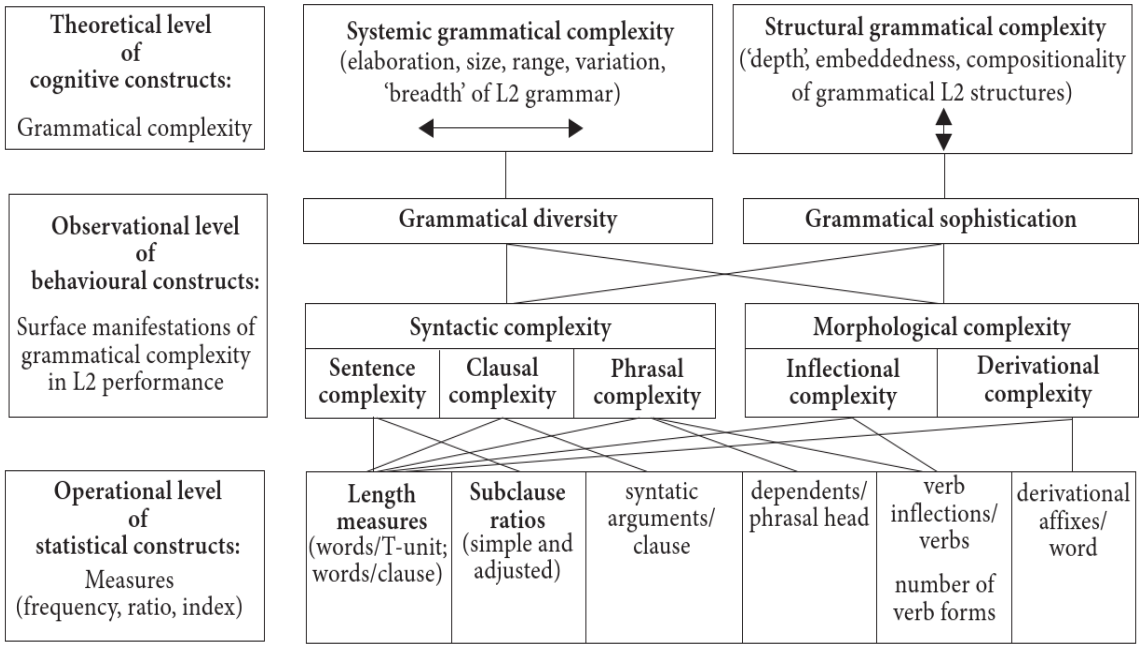


Figure 1. A taxonomy of complexity constructs

Bulté and Housen (2012, p. 23) argue that L2 complexity comprises relative and absolute complexity (see the top part of Figure 2.1). According to them, relative complexity is better called 'difficulty'. They describe it in this way: "A language feature that is costly, difficult, or difficult for some learners or users may be less costly, less difficult, or even easy for other learners or users, depending on such individuality factors as their level of SFL development, language aptitude, memory capacity, L1 background, and motivation." This theoretical framework serves as the foundation for this dissertation. In addition, Bulté and Housen said that beyond these subjective, learner-dependent elements related to difficulty, additional objective, learner-independent factors can influence the ease with which SFL features are learned and processed and the difficulty with which they are processed. Following Bulté and Housen's (2012), Figure 2.1 breakdown of L2 complexity, the current study focused on the functional complexity at the sentence, phrase, and clause level by exploring the individual syntactic complexity features that make up the learner's SFL writing system. I analysed syntactic complexity using the model of L2 complexity developed by Bulté and Housen (see Bulté and Housen 2012; Fig. 1, p. 27). In addition, they assert that SC is an element of functional complexity and structural complexity. Structural complexity is a component of linguistic complexity and absolute complexity. One of the two components of the SFL complexity is its absolute complexity.

In the current dissertation, Figure 2.2 illustrates how to investigate syntactic complexity, a subset of linguistic complexity. We have sentence, clause, and phrasal complexity at the observational level for L2 performance. Various syntactic complexity indices at the operational level can be used to examine syntactic complexity. Consequently, this study examined syntactic complexity as a branch of linguistic complexity at the observational level (see Figure 2.2 below and Bulté and Housen 2012; Fig. 2, p. 27) by employing over a thousand real-world language usage instances in the form of essays. This study gives information regarding the characteristics of syntactic difficulty in EFL writing at different competency levels (mainly answering the question of how the various CEFR proficiency levels differ in terms of SC). The operational level of this study (see Figure 2.2 and Bulté and Alex Housen 2012, Figure 2, p. 27) is based on analytical measurements that give us concretely quantified degrees, ratios, and frequencies of syntactic complexity in samples that are more objectively measured. The current study investigates systemic and structural grammatical complexity on a theoretical level, following which it investigates grammatical diversity and sophistication on an observational level of behavioural constructs.

Figure 2.2 L2 complexity at different levels of analysis (Bulté & Housen, 2012, p.27)



In addition to Bulté and Housen's (2012) description of individual factors contributing to SFL complexity in writing, the following factors may be associated with SFL complexity in writing: task design, repetition of the writing task (that means writing the same task again), task complexity, the topic, timed conditions, writing situation, mode of discourse, instructional factors, and age and grade level. Even more aspects may be related to the development of SFL complexity.

SFL complexity has been defined in much research at the operational level (see Figure 2.2), but only a few have acknowledged grammatical complexity in terms of behavioural components (Bulté et al., 2008; Bulté & Housen, 2012; Ortega, 2003; Norris & Ortega, 2009; Skehan, 2003; Skehan & Foster, 2005). Bulté and Housen (2012) argue that further research is needed to better understand grammatical complexity as a theoretical construct and the validity of complexity measures utilized in empirical research. This research focuses solely on complexity and its sub-branch syntactic complexity at the operational level. It excludes all other branches of complexity from consideration (Housen et al., 2012, Figure 2; Bulté & Housen, 2012; 2014).

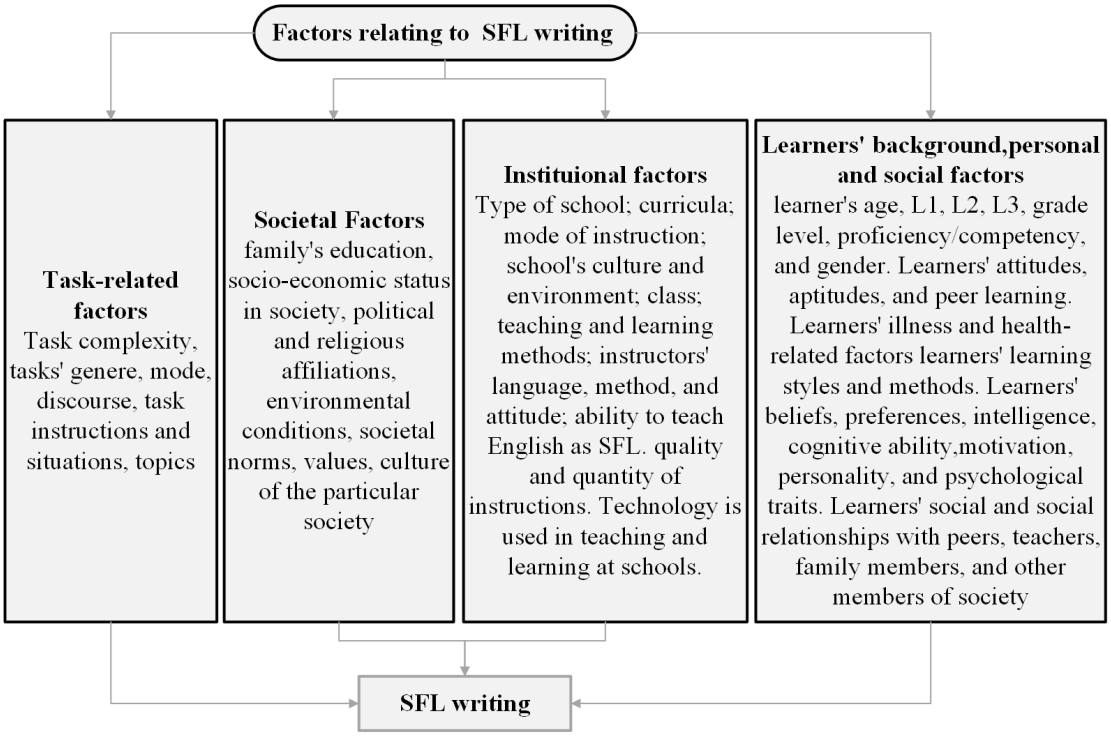
2.2.2 Factors relating to SFL writing

Some elements that may influence SFL writing are rarely acknowledged, investigated, or studied in SFL writing research. The current dissertation, for instance, explores the link between syntactic complexity, age, and L1, as both learners' age and L1 might affect syntactic complexity in their writing in the SFL. Many other things could also influence the evolution of SFL writing. Therefore, such aspects should be explored in sufficient depth in future research to comprehend better the linguistic basis of the CEFR levels and how they differ in terms of their linguistic features.

Next, some factors that could influence SFL writing are discussed (see Figure 2.3). Teaching and learning that consist mainly of planned activities and the repetition of similar writing tasks may be two characteristics of SFL teaching in some contexts. Numerous task-related aspects include increasing task complexity, using diverse themes, engaging in timed or untimed scenarios, and employing diverse forms of speech. Then there are situational, contextual, and environmental aspects, such as developing spontaneous and organised writing sessions and scenarios, learning using different instructional strategies, and learner-related factors, such as learning at an older age or grade level. This dissertation emphasises the significance of investigating all facets of SFL writing.

Identifying the elements that influence second-language writing is a crucial question. It is difficult to answer this question because only some studies can consider all the factors. The current dissertation examines syntactic complexity in SFL writing and a few variables that can influence writing in SFL, namely the learners' L1 and age/grade level. Many other elements can influence SFL writing, SFL proficiency, and its development more generally. A number of the factors that can affect SFL writing are summarised in Figure 2.3. Finally, it is also feasible that some factors influence SFL development and competence at a specific time (or at a particular proficiency level) more than at other times (or proficiency levels).

Figure 2.3 Factors that potentially affect SFL writing

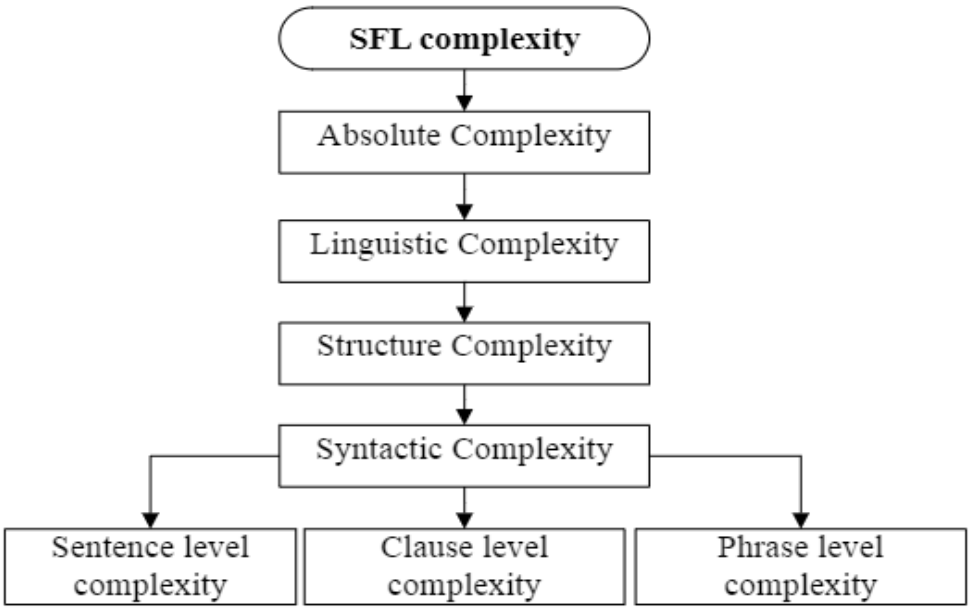


3 SYNTACTIC COMPLEXITY (SC)

Syntactic complexity is an essential feature of SFL students' and test-takers' language proficiency and the key theoretical concept of this dissertation. Investigating syntactic complexity and its relationship to more general language proficiency is essential in developing, acquiring, assessing, and teaching SFL learners' languages (Ortega, 2009; Bulté & Housen, 2014). Bulté and Housen's (2012, see Figure 2.2) paradigm posits the possibility of observing syntactic complexity at the observational level. It has various syntactic patterns and a spectrum of operational sophistication. Combining syntactic structures and constructions results in complex features observed in the learners' texts.

Language learners' language contains syntactic structures and levels of sophistication. Each piece of writing is unique in terms of the syntactic structures it employs and its sophistication. According to Lu (2011) and Ortega (2003), research in foreign language writing and education has emphasized the importance of syntactic complexity in learners' targeted language. I investigated syntactic complexity, a component of functional complexity, a component of structural complexity, a component of linguistic complexity, and a component of absolute complexity. Based on Bulté and Housen's model of L2 complexity, I simplified the model to present what I have studied in the current dissertation (see Figure 3.1)

Figure 3.1 Simplified model of syntactic complexity as a component of SFL complexity explored in the current dissertation at the sentence, clause, and phrase levels.



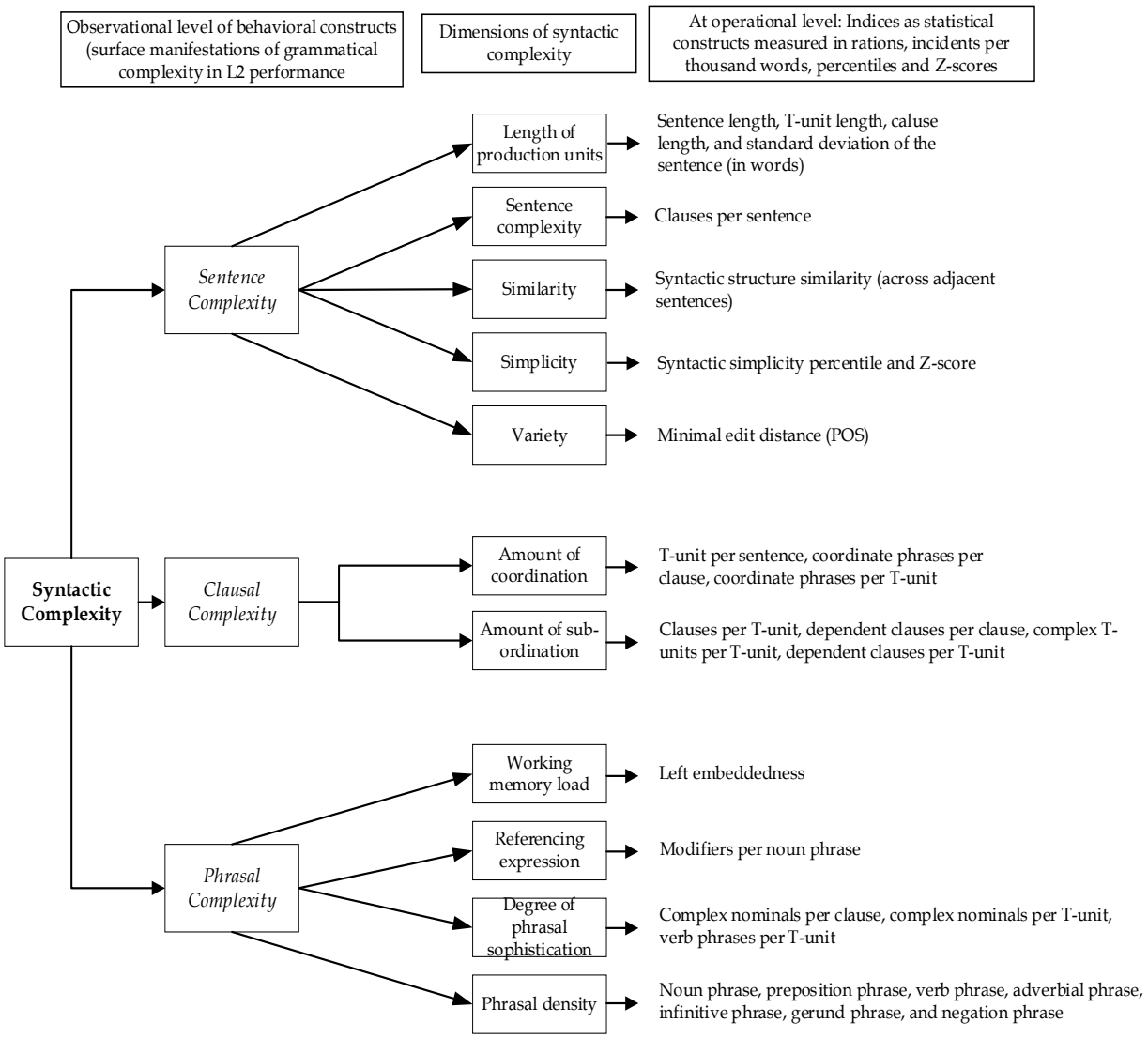
3.1 Definition of Syntactic Complexity

There are several definitions of syntactic complexity.

- "The range of forms that surface in language production and the degree of sophistication of such forms" Ortega (2003, p. 492)
- "Syntactic complexity, as a sub-component of linguistic complexity, is itself a multi-layered construct consisting of distinct sub-constructs that relate to different sources of complexity which each must be gauged by different measures" (Bulté & Housen 2012, p.40; based on views of Ortega 2003; Spoelman & Verspoor 2010).
- "Grammatical complexity is usually linked with elaboration and clausal embedding in linguistic theory" Biber et al., 2016. p. 87.

Biber et al. (2016) called syntactic complexity grammatical complexity. The current dissertation uses the following definition of SC that combines many elements of the definitions reported in the literature: ***Syntactic complexity is multifaceted, multidimensional, and multi-layered; it has a range and sophistication of forms and structures on both the surface and at deep language levels.***

Figure 3.2 Syntactic complexity explored in this dissertation at operational level



3.2 Syntactic Complexity in different fields

Syntactic complexity has been investigated in different fields, not only in applied linguistics and its branches like, for example, corpus linguistics and second language acquisition research. For instance, in recent years, there has been a growing body of research on syntactic aspects in marketing (e.g., Miller et al. 2015), health sciences (e.g., Bhat et al. 2015), and brain research, where Timmers et al. (2015) and Bhat et al. (2015) looked at how the brain plans syntactic complexity, Peelle et al. (2004) examined SC from the point of view of sentence understanding, and Caplan (2013) investigated syntax comprehension in aphasic language processing. Research on SC in speaking and L1 writing will be reported

next to provide some background to the research on syntactic complexity in SFL writing.

3.2.1 Syntactic complexity in speech

Most of the measurements used in syntactic complexity analysis in speech come from the writing realm (see the next section for details). These earlier measures can be similarly applied to voice transcripts for complexity computation. Several studies have observed the relationship between speech's syntactic complexity and the speakers' overall speaking ability (Halleck, 1995; Miles et al., 2001; Iwashita, 2008). Halleck (1995) found that three objective syntactic difficulty measures, including mean T-unit length, mean error-free T-unit length, and percentage of error-free T-units, were associated with holistic speaker evaluations.

Ratio-based measures are also widely employed in research on speaking, directly applying syntactic complexity measures in various contexts. In a study of Japanese EFL speakers by Iwashita (2008), length-based complexity indicators (i.e., the number of T-units and sentences per T-unit) were effective predictors of oral proficiency. The mean length of an utterance (Condouris et al., 2003), the word count or tree depth (Berninger et al., 2011), the mean length of T-units, and the mean number of clauses per T-unit are all further examples of the SC indices investigated in speaking research (Miles et al., 2001). Frequency-based measurements, such as the frequency of complete sentences in Berninger et al. (2011), were employed less frequently.

Speaking output could be tidier (e.g., considering disfluencies such as false starts, repetitions, and filled pauses). As a result, researchers may need to remove certain disfluencies before assessing the syntactic complexity features of speech. Furthermore, although automated speech recognition output lacks interpunctuation, the boundaries of phrases and sentences must be recognized for both their sentential and parser-based characteristics. As Chen et al. (2010) indicated, it is, however, possible to employ automated classifiers trained to predict clause and sentence boundaries.

3.3 Syntactic complexity in writing in English

Research on syntactic complexity in written English goes back to the late 19th century and started with analysing Old English (Anglo-Saxon) texts, Bible translations, and Shakespearean prose. Old English is the language the Anglo-Saxons used (from ca. 500 to 1066, when they ruled most of Britain). Smith (1893) looked at the sequence of words in sentences, phrases, and clauses and syntactic rules in the Anglo-Saxon language. He focused on the frequency of instances of the clauses in these writings. Smith (1904) appears to be the first to investigate the length of sentences, concluding that the unit of syntax is the breath group, which refers to a phrase or clause rather than a complete sentence. Smith (1904)

also paid particular attention to relative pronouns in analysing English syntax. Analyses of the English translations of the Bible over different centuries also contribute to research on syntax and its complexity (e.g., Grainger, 1907).

3.3.1 Syntactic Complexity in English as a first language (L1)

This chapter gives a brief outline and history of research on syntactic complexity in English as the first language (mother tongue, or L1). The chapter traces such research from the early twentieth century.

O'Shea (1907) investigated the linguistic development of English-speaking students and noted that when reading, the students evaluated separate components of the sentence and got familiar with the single words as they progressed. Rather than saying each word individually, students created a sequence demonstrating their understanding of the syntactical relationships and meaning of the text. O'Shea (1907) also noted that when learning a new language, the students could not comprehend what they were reading unless they understood how the words worked together. Neilson (1908) attempted to determine the reasons for the student's failure to attend school and concluded that their linguistic development, particularly their ability to produce long (and, thus, complex) sentences, was the primary reason for the student's inability to attend school.

Some researchers were interested quite early in whether age is associated with syntactic complexity. LaBrant (1932) introduced the concept of "clause" as a necessary measure for investigating syntactic complexity. The studies were motivated by the desire to determine the number of subordinate clauses (LaBrant, 1932), the number of sentences in texts, the complexity of sentences (Bear, 1939; Hoppes, 1934) and their relationship with writers' age. According to the findings of these studies, the number of subordinate clauses, the number of sentences, and the complexity of the sentences all increased with age. Older children and more experienced adults wrote longer and more complex sentences than younger students. When mastering a language, age and experience appear to be critical factors. More recent research, mainly in the 1960s and 1970s, based on sentence length and T-unit length, has discovered that these SC features develop with learners' age (or grade level) (Hunt, 1965, 1966, 1968, 1969, 1970, Hunt and O' Donnell, 1970). Particularly Hunt's (1965) introduction of the T-unit (defined as the "minimally terminable unit") has been influential for research on syntactic complexity as it has become one of the most often investigated SC features. Researchers like Cooper (1976), Monroe (1975), and O'Donnel (1967, 1968) have argued that the T-unit is an excellent way to measure how complex learners' syntax is.

4 SYNTACTIC COMPLEXITY WITH REFERENCE TO NON-CEFR PROFICIENCY LEVELS

The following two main chapters, four and five, summarise prior studies on syntactic complexity in English as a second or foreign language related to learners' proficiency. More specifically, chapter four examines studies in which the learners' proficiency level is characterized by their educational attainment, a generic stage of language learning (e.g., beginners), or a specific proficiency framework or scale (e.g., IELTS bands). All these ways of determining learners' competency have one thing in common: they do not refer to the CEFR levels, which served as the foundation for the research documented in this dissertation and whose conclusions are thus not directly or easily comparable to the present author's findings. In comparison, chapter five discusses research on SC in SFL English that employed the CEFR levels as a benchmark for their informants' skills and whose conclusions are more directly comparable to those in the current dissertation.

Proficiency in second and foreign languages was measured or determined in the non-CEFR studies using a range of criteria and methods. Most research indicates a considerable correlation between syntactic complexity and proficiency. However, the different findings regarding SC could be easier to interpret, as proficiency has been operationalized in many ways using different methods and criteria. The most frequently utilized standards include:

- program or grade levels
- length of language studies
- holistic or analytical ratings using different rating scales
- scores on language tests measuring a range of skills (not writing or not only writing)

The following sub-chapters present research on syntactic complexity from longitudinal, task-based, and cross-sectional studies.

4.1 Syntactic complexity in longitudinal and cross-sectional studies

4.1.1 Syntactic complexity features studied in longitudinal studies

This section summarises longitudinal research, particularly studies that explored the relationship between syntactic complexity and non-CEFR proficiency levels. This sub-chapter analyses the findings, including participants with L1 and SFL levels and individuals of varied ages and credentials (see Appendix Tables 1 and 2 for details of these studies). The emphasis here is on syntactic complexity characteristics.

This section describes research that used samples of English-language texts produced by English learners. Participants wrote two or more texts separated by a time gap. Several studies have been conducted on syntactic complexity, particularly on the T-unit. For instance, Ishikawa (1995) examined two essays written three months apart by novices. Similarly, Larsen-Freeman (1983; Study 3) assigned five essays at two-week intervals to upper-intermediate learners. Both studies discovered significant variations in T-unit length between the texts. The T-unit length fluctuated arbitrarily and without pattern. Casanave (1994) discovered relatively significant differences in the length of T-units, the number of phrases per T-unit, and the number of complex T-units per T-unit among the texts. In contrast, Arthur (1979), Ishikawa (1995, Analysis 1), and Tomita (1990, Level 1: Sophomores, Level 3: Seniors) found that T-unit length did not differ statistically between texts created at different times. Ishikawa (1995) used writing samples from group number 2, a beginning class that wrote two essays three months apart; the results suggested that the number of clauses per sentence increased considerably between the two samples.

Storch (2007) evaluated argumentative essays written a semester apart and discovered no statistically significant difference in the averages of clauses per T-unit and dependent clauses per clause measured at one versus time two. Raquel Serrano et al. (2012) analyzed descriptive essays written three times over a year and identified significant differences in the length of T-units between Time 1 and Time 3. Similarly, they discovered a statistically significant rise in phrases per T-unit between times 1 and 3 and between times 2 and 3. There were, however, no significant differences in the length of T-units between Time 1 and Time 2 or between Time 2 and Time 3, and no significant differences in the number of clauses per T-unit between Time 1 and Time 2.

Mazgutova and Kormos (2015) recently demonstrated statistically significant results in the syntactic structural similarity of two tasks written at one-month intervals (in their Group 1). Additionally, after one month, they demonstrated a statistically significant decrease in syntactic structural similarity but increased modifiers per noun phrase and complex nominals. In Group 2, significant increases were found for conditional, relative, and sophisticated post-modifiers when comparing two assignments written one month apart. Finally,

when two exams written one month apart by Group 1 were compared, they discovered a considerable rise in infinitive clauses. Mazgutova & Kormos (2015), on the other hand, discovered no statistically significant differences in conditional clauses, prepositional phrases, relative clauses, simple post-modifiers, complex post-modifiers, mean length of T-unit, dependent clauses per T-unit, modifiers per noun phrase, and complex nominals between two tasks written separately by group 1 after one month. Between two assignments one month apart in Group 2, there were no significant differences in prepositional phrases, infinitive clauses, simple post-modifiers, T-unit mean length or dependent clauses per T-unit.

The most current study in this subchapter, Rosmawati (2019), analyzed academic written essays over one year at twenty measurement points, utilizing twenty essays on various topics. Rosmawati noticed a general rise in the complexity of syntactic structures, including mean clauses and sentence length. Because findings are presented at approximately 20 points for each participant, the two indices did not grow or decline linearly. The researcher made no distinction between consecutive and non-consecutive essays, meaning there were probable cases in which significant differences between different time points could have occurred.

The **conclusions** that can be drawn from these longitudinal studies on SFL English are that many studies have found increases in SC over time, which is in line with other research and theorizations that argue that SC increases as SFL proficiency grows. Some studies have shown no change over time, however. This can be due to several reasons. For example, the interval between the data gathering time points was too short to allow changes to occur in the particular SC indices investigated. Related to this, the indices in question may stay the same over time, or if they do, it takes much longer to happen than what was covered in the research. Also, task-related factors may have affected the results, as the task(s) used were not sensitive enough to reveal changes in the examined SC indices. Finally, as the studies did not refer to the CEFR levels, it is difficult to relate their findings to the current dissertation's findings.

The review provides evidence that many SC indices are likely to change as learners' proficiency improves, so it can be expected that there might be at least some apparent differences between learners whose CEFR levels differ. This is all the more likely since the CEFR levels are broad; they represent different stages of proficiency (even if it is challenging to determine where precisely one level ends and the next starts). Moving from one CEFR level to the next takes considerable time, allowing changes to emerge in the learners' use of different linguistic devices, including syntactic features.

4.1.2 Syntactic complexity explored in cross-sectional studies

Cross-sectional task-based studies: Task-based teaching and learning research has also investigated syntactic complexity among learners of SFL English. Since the beginning of the new millennium, studies have concentrated on task complexity (Robinson, 2002), cognitive complexity (Robinson, 2005), task-based language

learning (Nunan, 2005), planning and task performance characteristics (Ellis & Yoon, 2005), and task-based language research (Ellis & Yoon, 2005). One set of research looked at task complexity and writing tasks. Kuiken et al. (2005) studied the impact of task complexity in two experimental tasks (a letter and a descriptive-cumulative argumentative). Across low and high-complexity tasks, they found no statistically significant variations in sentences per T-unit, dependent clauses per clause, or clauses per T-unit. Another study by Michel, Kuiken & Vedder (2007) compared cognitive task complexity between more and less sophisticated task types. They found no significant differences in clauses per T-unit or dependent clauses pe clause between first- and third-year students.

Cross-sectional studies in general: This section summarises the studies (details in Appendix 1, Tables 1 and 2) that examined syntactic complexity concerning non-CEFR proficiency levels. Proficiency levels were determined in a variety of ways in these studies. Certain studies quantified proficiency using programme levels, while others referred to test scores, holistic ratings, and school levels as indicators of learners' ability.

When syntactic complexity began to be investigated more systematically in the 1970s and 1980s, research focused on the T-unit in language learners' writing, particularly the mean length of the T-unit. Hunt (1965) argued for the importance of the T-unit in studying syntax and language development. He defined the T-unit as "one main clause plus the subordinate clauses attached to or embedded within it" (Hunt, 1965, p. 49). However, some other length measures were also explored in these early SC studies, as described in more detail below (see Appendix 1, Tables 1 and 2). These measures included the mean length of clauses and sentences (in words). In SC studies conducted in the 1970s, 1980s, and 1990s, length was not the only SC measurement employed. One was the number of clauses per T-unit (C/T), which is a length measurement and indicates another T-unit level or type of complexity (see, e.g., Monroe 1975; Flahive & Snow, 1980). Some studies investigated the number of dependent clauses per T-unit or per clause (e.g., Vann, 1979; Homburg, 1984; Hirano, 1991), representing an apparent expansion in the studied SC indices.

What was typical of SC research before the 2000s was the relatively small number of analyzed learners and learner performances. Most of the studies included only a few dozen learners and rarely managed to collect data from more than 100 learners (see Appendix 1 and Tables 1 and 2 for details). The apparent reason for this was that most of the data had to be manually coded (annotated), which could be faster but more laborious. An evident quantitative and qualitative change in research on SC, particularly in English language learners' writing, can be seen after the millennium, particularly in the 2010s. First, this recent research is more large-scale: most of the studies reported in the literature include over 100 learners or texts, often several hundred. Besides this quantitative difference in the data, the studies have changed qualitatively and typically investigated a more comprehensive range of SC indices. Subordination and coordination indices, particularly phrasal level indices, are now being investigated regularly in

addition to the more traditional indices relating to the length of T-units, clauses, and sentences.

The main reason for this change is the introduction in research studies of automated tools for analyzing texts written in English, particularly the L2 Syntactic Complexity Analyzer and Coh-Metrix. Most studies mentioned in Appendix 1 and Tables 1 and 2 (regarding CEFR levels) used one of these automated tools to analyze the syntax of their English language learners' writing. Sometimes the writing of native English speakers is also tested if the study involves comparing L1 and English as a foreign or second language speaker.

Below, some early SC studies from the 1970s to 1990s are described, followed by a similar description of the more recent studies from the 2010s. All these studies referred to something other than the CEFR levels to define the participants' proficiency in English (i.e., they used learners' course level, years of study, or language test scores as proficiency indicators). The studies described below are selected to illustrate what was typically investigated in SC research at the time. The focus is on the larger-scale studies (particularly when earlier research is described); a more detailed picture of this research can be obtained in Appendix 1, Tables 1 and 2.

Gipps and Ewen (1974) investigated the first 100 words of essays written by 751 school-aged children who spoke various first languages and English as a second language; the children wrote on one of three provided topics. Years of L2 education were taken into consideration for determining proficiency. They discovered that the length of T-units differed significantly depending on the number of years of L2 schooling. Larsen and Freeman (1978) investigated 212 learners and used five programme levels to assign proficiency levels to their informants. The students were required to write 30-minute essays on a single topic. The length of the T-unit varied significantly among programme levels. Flahive and Snow (1980) investigated 50-minute essays on one of several assigned topics produced by 300 English language learners divided into six program levels; holistic ratings based on a 5-point scale were carried out at each of the six programme levels. They discovered that the length of T-units and the number of clauses per T-unit varied significantly across skill levels. Tomita (1990) analyzed the descriptions of visuals written by 258 high school students. Students were drawn from three different school years and divided into three groups based on their grades (low, mid, and high). Tomita discovered that the length of T-units differed significantly between people with low vs high holistic ratings. Ferris (1994) looked at 160 students entering university with a 1–10 holistic rating and then divided them into two groups: 60 low-level students and 100 advanced-level students. The discriminant analysis results revealed a substantial difference between low- and high-level learners in the number of words, prepositional phrases, negatives, adverbs, first and second person, and impersonal pronouns.

McNamara et al. (2010) investigated 120 untimed essays on four different topics written outside the classroom. Citing outside sources was permitted in this case. There was an unequal number of pupils for each prompt. There were

statistically significant differences between high and low-proficiency writers in left embeddedness but no differences in the number of words per phrase. Crossley et al. (2011) investigated 120 untimed argumentative essays evaluated using the SAT's holistic rubric. Modifiers per noun phrase showed statistically significant changes between 9th graders, 11th graders, and first-year college students. There were statistically significant differences also in left embeddedness between learners with low and high competence levels.

Guo et al. (2013) investigated 240 integrated and independent essays written in 30 minutes. The essays were graded using independent and integrated writing rubrics used in the TOEFL iBT examination, with skill levels ranging from 1 to 5. The outcomes of the integrated essays were statistically significant. Past participle verbs, verbs in base form, personal pronouns, verbs in the third person singular present form, verbs in the past tense, modifiers per noun phrase, the to-infinitive, and prepositional phrases yielded statistically significant correlations between the scores and the above-mentioned linguistic features. Independent essays also yielded statistically significant results for the following variables: past participle verbs, verbs in non-3rd person present singular form, embeddedness clauses, number of modifiers per noun phrase, personal pronouns, and base form verbs. Crossley et al. (2014) examined 480 independent essays scored according to the TOEFL iBT independent writing rubrics 1 through 5. Significant relationships were found between human ratings and noun phrase density, modifiers per noun phrase, agentless passives, perfect verb forms, the incidence of conjuncts, determiners, minimal edit distance, and adjective phrases.

Thilagha Jagaiah (2017) examined argumentative, informative, explanatory, and narrative essays by 1029 students. The pupils were in the eighth grade and came from various backgrounds. Students were rated using an automated essay scoring (AES) engine called Project Essay Grade (Page, 2003), as well as six sub-scores based on a total of six sub-scores based on writing quality (Page, 2003). Students were given six sub-scores depending on their writing quality: overall growth, organization, support, sentence structure, word choice, and mechanics. The first two proficiency bands assigned to the students included 115 students; the third, fourth, and fifth bands included 914 students. Two categories were established: at-risk (basic bands 1 and 2) and not-at-risk (bands 3–5; proficient, advanced learners). The essays were analyzed for noun phrases, adjectives, agentless passive voice (density occurrence), and other noun phrases. Also analyzed were the number of words (sentence length), the frequency of temporal connectives, the frequency of all connectives, the frequency of causal connectives, the number of modifiers per noun phrase, the minimum edit distance, the part of speech, the minimum edit distance for all words, and the number of modifiers per noun phrase. It was demonstrated that the difference between the writings of at-risk and not-at-risk learners was statistically significant for all these indices.

Kyle (2016) investigated two argumentative essays (two topics / prompts) written by 240 English language learners; the data set was drawn from the TOEFL test essays and judged holistically on the 5-point TOEFL score scale. The researcher studied the extent to which various tradional indices of SC (e.g., those

included in the L2 Syntactic Complexity Analyzer) and more fine-grained indices of phrasal complexity and sophistication could predict the holistic TOEFL writing scores. These latter indices differ from the larger scale absolute indices of complexity in that they are based on frequency and usage. Kyle found that the L2SCA indices that could be used (mean clause length, coordinate phrases per clause, and complex nominals per T-unit) could explain a significant but modest amount of variance in the scores: mean clause length explained 4% of variance (the amount of variance varied from about 6% for essay topic 1 to about 9% for topic 2). The main finding of Kyle's study was, however, that the more fine-grained phrasal level indices could explain between 18 and 25% of variance in the learners' holistic writing scores, that it, they were better predictors of writing than the more traditional indices. (See also Kyle and Crossley (2017, 2018) that are based on the data gathered for Kyle (2016)).

The **main takeaway** from the numerous studies on SC based on various ways to determine learners' proficiency levels is that in most of the studies, there were significant differences between learners at different proficiency levels and that **syntactic complexity tended to increase with proficiency**. These findings also suggest that similar results can be expected for at least some SC indices when learners' proficiency is determined systematically using the CEFR levels. Another takeaway (for future research) is that phrasal level, more fine-grained indices may be better predictors of SFL writing and, thus, may offer even better ways to distinguish CEFR levels than the traditional larger scale indices of SC.

With some exceptions, the longitudinal and cross-sectional studies reviewed above have been limited regarding the number of SC indices investigated, the size of the corpus of writing samples, or both. The introduction of automated analysis tools has made it possible to include more SC indices in recent studies. Thus, more research is needed to obtain a more comprehensive view of syntactic complexity in SFL writing by investigating it as a multidimensional phenomenon with a relatively large number of SC indices. In addition, the fact that the studies reviewed above used a range of approaches to defining learners' proficiency makes comparing their findings difficult. Therefore, defining learner proficiency concerning such a widely used framework as the CEFR offers a way to obtain results that are potentially easier to relate to other studies using the same framework.

The next chapter will examine the research on syntactic complexity in English as a foreign or second language writing that has been done with reference to the Common European Framework of Reference for Languages (CEFR) levels.

5 SYNTACTIC COMPLEXITY IN SFL ENGLISH WRITING WITH REFERENCE TO THE CEFR LEVELS

This chapter introduces the CEFR and how syntactic complexity is defined in it. The section then focuses on studies that investigated syntactic complexity in relation to CEFR levels and summarises these investigations. This chapter will synthesize the literature on CEFR-related studies from 2001, when the CFER was introduced, until the present.

5.1 The CEFR and its historical perspectives (an introduction)

The Common European Framework is a common framework for developing language syllabuses, curriculum guidelines, tests, textbooks, and other educational materials throughout Europe. Language learners must learn to do certain things to communicate effectively in a foreign language, and they must also acquire specific information and abilities to do so. The CEFR describes a wide range of such factors. The descriptions in the CEFR also include information about the cultural setting in which the language is used and learned. The Framework also lays out different levels of proficiency, which can be used to describe how well people master a foreign language at each stage of learning and throughout their lives (CoE, 2001, p. 1).

The establishment of the Council of Europe in 1949 was to extend and protect human rights among member countries. The Council signed the convention on cultural cooperation in 1954, and the Council also took responsibility for matters related to languages (CoE, 2001). Because of the Second World War, language learning suffered a significant setback. People's social ties were broken due to distrust and ignorance. Reconstruction and reformation took place in Europe during the years after the Second World War; for example, travel and communication infrastructures were rapidly evolving simultaneously. The

Council of Europe saw knowledge of other countries' languages as crucial for increasing mutual understanding between countries and their peoples. Therefore, promoting language education has been one of the activities of the Council since the very beginning.

5.2 The advent of the CEFR

The Swiss government proposed that the Council of Europe develop a Common European Framework for Languages (CEFR) at an international conference held in Switzerland in 1991. Other European countries attending the conference accepted the proposal (CoE, 2001, p. 5; the origins of the CEFR are discussed in CoE 2020 and North 2020 and 2021). Initially, the CEFR was established to discuss and describe what is involved in using and learning a foreign language. It includes general and linguistic competencies, activities, functions, tasks, and texts. A key objective of the Framework is to increase transparency and comparability among different language qualifications, curricula, and materials. It is based on extensive previous research and theorizing on action-oriented and functional language use (see North, 2020; 2021, for details); it considers language users and learners primarily as 'social agents', i.e., members of society who have tasks to accomplish within a given set of circumstances, within a specific setting, and within a specific domain of action (CoE, 2001). The revised CEFR (CoE, 2020; CEFR Companion Volume) extends the Framework's theoretical foundation by emphasizing language users' plurilingual competence and mediation (using and switching between different languages). Even though the original CEFR of 2001 had the most significant impact on language assessment, the updated CEFR CV aims to directly impact language teacher education and classroom instruction (North, 2021).

The selection of the six levels for the CEFR was considered sufficient to distinguish between various degrees of proficiency. It was felt that these six levels would provide enough detail and granularity to accurately measure a learner's proficiency level, while still providing a clear and concise framework for language learners and teachers. However, many scales in the CEFR split some of the primary levels into two (i.e., into a lower and higher sub-level), particularly levels A2, B1, and B2). This was done to reflect the vast range of language proficiency more accurately and to provide a more nuanced understanding of an individual's language skills. Additionally, it provides a more detailed view of the language proficiency levels, which makes it easier for language teachers and learners to accurately measure their progress. A "can-do" strategy was used to identify broad and specific activities and skills regarding language proficiency. Notably, although the CEFR has scales that describe linguistic competence, they need to be more detailed. This is where the "can-do" strategy comes in, to ensure that language proficiency levels are more accurately measured and properly described. They must define what lexical and grammatical features are typical of particular languages at particular proficiency levels. The "can-do" strategy

focuses on identifying the language use proficiency of learners in terms of what they can do with language, rather than what they know about it. This means that it takes into account the various language skills and the specific language features that learners need to be able to use accurately and effectively in order to reach a certain level of proficiency. This is how the CEFR is able to provide more detailed descriptions of language proficiency levels. The CEFR is intended to apply to all languages. Therefore, the scales define what learners can do or, in some cases, what kind of linguistic competence is expected at each level (see the next section). The CEFR divides language proficiency into 6 levels: A1, A2, B1, B2, C1, and C2. Each level is further divided into two sub-levels, allowing the user to gain a more nuanced understanding of their proficiency level. The CEFR scales are applicable across all languages, making it a useful tool for assessing language proficiency in a variety of contexts.

5.3 Syntactic complexity in the CEFR

The syntax and syntactic complexity are acknowledged explicitly in a few places in the Common European Framework of Reference for Languages. The CEFR defines syntax as follows: "Syntax deals with the organization of words into sentences in terms of the categories, elements, classes, structures, processes, and relations involved, often presented in a set of rules," and it states that the syntax of the language of the learners is "highly complex and largely unconscious" (CoE, 2001, p. 115). Additionally, the CEFR descriptors for the B2, C1 and C2 levels refer to "complexity" (Table 5.1). On page 110 of the CEFR, the term "complex" is used to refer to complexity in the general linguistic range scale, which states that a B2 level learner "has a variety of language to give clear descriptions, express viewpoints, and develop arguments without much conspicuous searching for words, while using some complex sentence forms to do so."

Table 5.1 References to complexity in scales of the CEFR (2001)

the CEFR Level	References
C2	clear, complex, logical structure (p.61), a wide range of language (p.110), complex language (p. 28, 114), complex reports, articles, or essays (p.62)
C1	clear, well-structured texts of complex subjects (p.61), appropriate formulation (p.110), a broad range of language (p.110), clear, well-structured expositions of complex subjects (p.62)
B2	clear, detailed texts on a variety of subjects (p.61), longer complex stretches of speech (p.129), some complex sentence forms (p.110)
B1	straightforward ...texts (p.61), a repertoire of frequently used patterns (p.114), expressing more complex thoughts (p.112), simple face-to-face conversation (p. 86,124), can write short, simple essays, brief reports (p.62)
A2	simple phrases and sentences (p.61), simple face-to-face conversation (p.86,124), basic language (p.110, 122), simple structures (p.29, 114), simple sentences (p.125), basic sentence patterns (p. 29, 110)
A1	simple isolated phrases and sentences, (p.61), basic range of simple expressions" (p.110), simple sentence patterns (p.114)

The terms representing complexity are often mentioned in and out of the descriptors/scales in the CEFR (CoE, 2001) and the CEFR companion volume (CoE, 2020). Complexity for A1 and A2, proficiency level learners is "simple" and "basic" on all scales. The term "complex" is used at B1 and higher proficiency levels. The term "complex" is used forty-six times in the B2, C1, and C2 proficiency scales, suggesting that complexity starts at the B2 proficiency level in the CEFR. Besides, the ALTE scales in the appendices to the CEFR (CoE, 2001, pp. 251-254) use the term "complex" seven times at higher levels, and the DIALANG scales (CoE, 2001, pp. 231-250) twenty-two times. In contrast, the word "simple" is used thirty-four times, and "basic" is used forty-six times at the A1 and A2 proficiency levels, which suggests the learners' language is not very complex at these levels, according to the CEFR.

Table 5.2 Number of terms related to complexity used in the CEFR (CoE, 2001)

Term	Complex	Simple	Basic	Syntax	Syntactic
Times	118	280	78	8	7

Examining the phrases linked with complexity makes it possible to discover additional information about how the CEFR presents complexity. First, the complexity-related terms in Table 5.2 include syntactic complexity; morpho-syntactic complexity, syntactic knowledge; ambiguity; errors; precision; oversimplification, and syntactic features. Thus, Table 5.2 is about the words associated with syntactic complexity.

The B1 level is transitional between simple and more complex language since terms referring to simple and rudimentary structures are used at levels below it and terms such as complex above it. Using the categories of basic, simple, and complex, we may describe language growth from simple to complex as a person's ability level increases, with B1 functioning as a bridge between simple and complex communication competencies. This analysis of the words used in the Common European Framework of Reference for Languages suggests that, even though the CEFR does not describe linguistic development in detail, the terminological choices in its descriptor scales suggest an underlying assumption of development from simple to more complex language (presumably including syntactic complexity). At the same time, learners' abilities to express more complex content and ideas also develop.

Next, I will further investigate how the CEFR refers to syntactic complexity. Complexity and its opposite, simplicity, are discussed in the Framework (see Tables 5.1 and 5.2), but the syntactic difficulty is rarely specified explicitly. There are frequent inconsistencies or absences of explicit references to the degree, frequency, and ratios of syntactic complexity structures across the CEFR levels. A CEFR A2 student "can write a succession of simple phrases and sentences connected using simple connectors" (CoE, 2001, p. 61). This connects to the vague approach's grammatical complexity, which begs the question: what do "simple words and sentences" refer to in this context? For instance, the scale does not explain what constitutes an average A1 in terms of syntactic simplicity or where the line between A1 and A2 is drawn for this linguistic feature.

Syntactically challenging texts, such as those specified in the upper CEFR levels, are likely to be longer and more complex in their content (e.g., reports and essays). They can contain numerous syntactically complex statements, although they are not explicitly specified on a particular scale. In addition, according to the CEFR, "particularly complex syntax consumes attentional resources that would otherwise be available for dealing with content; for example, long sentences containing a large number of subordinate clauses, non-continuous constituents, multiple negations, scope ambiguity, and the use of anaphoric and deictic without clear antecedents and references" (CoE, 2001, p.165). Some references to linguistic complexity, which could be interpreted as referring to syntactic difficulty, are also included in the comprehension scales. However, these are not particularly related to syntactic complexity (e.g., overall listening comprehension, p. 66, listening as a live audience member, p. 67). In addition, the CEFR addresses the complexity of topics, issues, subjects, information, directions, exchanges, and argumentative lines. However, the CEFR references to text complexity primarily

concern the text's content and structure rather than its grammatical/syntactic patterns.

An essential empirical concern regarding syntactic complexity is how "simple, basic, and complex" syntax relates to the CEFR levels, that is, to language learners' (writing) skills as reflected by the CEFR levels and (as is the case in this study) if and how syntactic complexity differs between different CEFR levels. A related question is whether syntactic complexity changes across L1 speakers of the same target language whose CEFR writing level in that target language is the same (for example, learners of English as a foreign language by Sindhi L1 speakers in Pakistan and by Finnish L1 speakers in Finland), or whether the syntactic complexity differs between learners of different ages even though their overall CEFR writing level is the same.

5.4 Syntactic complexity explored concerning the CEFR

This sub-chapter aims to summarise the research on the syntactic complexity of English language learners at different CEFR levels. There have been numerous calls by both second language acquisition (SLA) and language testing researchers for research to investigate links between the CEFR and the findings from studies on language learning (e.g., Neff-van, 2013; Chen et al., 2014; Gyllstad et al., 2014; Hismanoglu, 2013; Hulstijn, Alderson and Schoonen, 2010; Hulstijn, 2007; Littlemore et al., 2014).

In the first empirical study on SC concerning the CEFR levels, Kim (2004) explored 33 written scripts by 33 Chinese university students learning English for specific purposes. The scripts were evaluated with reference to the CEFR levels A2, B1, and B2. Kim used indices based on clauses and T-units. She investigated different types of clauses: (1) adverbial clauses per clause, dependent clauses per clause, and adjective clauses per clause, (2) amount of subordination: clauses per T-unit, dependent clauses per T-unit, and phrasal sophistication: prepositional phrases per clause, participial phrases per clause, infinitive phrases per clause, and dependent clauses per clause. Kim reported apparent differences between A2 and B2 but non-significant and minor differences between the adjacent levels A2 and B1. She did not find differences between adjacent and non-adjacent levels in gerund phrases per clause or nominal clauses per clause (see Table 5.3 for a summary of the CEFR-related studies covered in this section).

Hawkins & Filipovic (2012) and Green (2012) investigated scripts from the Cambridge Learner Corpus (CLC). The CLC contains Cambridge English language exam responses on various topics, and many scripts have been analyzed in these studies. Hawkins & Filipovic (2012) and Green (2012) used the data of only those learners who passed with grades A, B, and C (and excluded failed participants) and were converted to CEFR levels A2 to C2 in the English profile studies. Hawkins & Filipovic (2012) reported significant differences in sentence length between A2 vs. B1, B1 vs. B2, B2 vs. C1, and C1 vs. C2, that is,

between every adjacent pair of CEFR levels covered by their data. Similarly, Green (2012) reported significant results in sentence length between B2 and C1, sentence syntax similarity (adjacent sentences & across all paragraphs) between B2 and C1, and noun phrase incidence between B1 and B2. However, Green (2012) found non-significant results for the other CEFR level pairs for the SC indices he investigated. Interestingly, unlike Hawkins and Filipovic's (2012) study, in Green's investigation, sentence length was not significantly different between levels except for B2 vs. C1. Also, Banerjee et al. (2015) examined essays written by students in response to two questions in the context of Cambridge examinations. The scripts were graded on a five-point scale of proficiency (A, B, C=C2=Pass), (D, E=below C2=Fail), based on holistic essay scores, pass, and fail in the Certificate of Proficiency in English (CPE) that targets C2 level. They found modifiers per noun phrase to differentiate between C2 and the levels below significantly.

Verspoor et al. (2012) explored descriptive texts written by 437 Dutch EFL learners from a pre-university secondary school program (12–15 years old). The texts were rated at proficiency levels 1 to 5 (beginner to advanced), linked to the CEFR levels in the following way: 1 = A1.1, 2 = A1.2, 3 = A2, 4 = B1.1, and 5 = B1.2. The study found T-unit length to be a moderate discriminator between the proficiency levels. They also found the proportion of dependent clauses and the proportion of simple versus complex sentences to be suitable separators of the levels. The relative clauses increased across all levels, but most clearly between A2 and B1.1. Similarly, the verb phrases used in different tenses were investigated, and the results showed that present and past tenses discriminated across proficiency levels 1-4.

Qi and Dong (2014) investigated 2600 argumentative essays collected from native and non-native learners of English. The learners' TOEIC scores were converted to CEFR levels. The study found that English as a foreign language learners' mean sentence length, clauses per sentence, and dependent clauses per sentence at the B1.2 level were significantly longer than those of English as second language learners. They found that mean sentence length and clauses per sentence differed between A2 versus B1.1, B1.1 versus B1.2, and B1.2 versus B2) in both EFL and ESL groups.

Gyllstad et al. (2014) explored 54 Swedish learners of English who were at three different levels in the Swedish school system. The 104 emails and stories they wrote were rated at A1-B1. The researchers merged the beginner's levels (A1 and A2) into one A level. The study found significant correlations between T-units, the number of words per clause, the number of clauses per T-unit, and the CEFR levels. They also found CEFR A levels and B1 levels to differ in mean T-unit length, clause length, and the number of clauses per T-unit.

In addition to the research listed above, Yoon (2017) examined argumentative essays authored by college-level Chinese EFL students on two distinct topics (40–80 minutes) and assessed them using TOEIC scores converted to CEFR levels. In detail, T-unit length, clause length, and complex nominals per clause showed significant differences between A2 and B2 level learners with

small effect sizes. However, between A2 and B2, they showed significant differences with large effect sizes.

Lahuerta Martnez (2018) investigated 188 Spanish L1 learners of English who wrote argumentative texts rated holistically on a 0–5 scale, converted to low intermediate (CEFR A2) and upper-intermediate (CEFR B1) levels based on the Oxford Placement test. The author reported high correlations between global scores and syntactic complexity measures (mean sentence length, noun phrase per clause, and complex sentence ratio). Similarly, the study reported moderate correlations between global score and syntactic complexity measures (compound sentence ratio, complex sentence ratio, compound-complex sentence ratio, coordinate clause ratio, and dependent clause ratio). In addition, the study reported significant differences between third- and fourth-year students in syntactic complexity indices (mean length of sentence, compound sentence ratio, complex sentence ratio, coordinate clause ratio, dependent clause ratio, and noun phrase per clause), which suggests that these SC indices separated A2 from B1 (assuming that most third-year students were at A2 and most fourth-year students at B1).

Furthermore, Lahuerta Martinez (2018) investigated whether the syntactic complexity is affected when the writers are of different genders. The findings indicated that the girls wrote more complex syntax than the boys. Boys and girls differed significantly in the mean length of sentences, compound-complex sentence ratio, coordinate clause ratio, and dependent clause ratio. Similarly, the study found significant differences between third-year boys and girls in the mean length of sentences and coordinate clause ratio. Besides, the fourth-year boys and girls significantly differed in the mean length of sentences, compound-complex sentence ratio, and coordinate clause ratio.

Paquot (2019) examined 78 university students' research papers submitted outside the classroom. The learners were undergraduate and graduate students with varying L1s. The papers were written on various subjects and scored with reference to the CEFR levels B2, C1, and C2. The results indicated that clauses per T-unit, dependent clauses per T-unit, dependent clauses per clause, clause length, verb phrases per T-unit, complex nominals per T-unit, and complex nominals per clause significantly differed across and within groups (B2 vs. C1, C1 vs. C2).

Finally, Polat (2019) examined argumentative essays submitted by 284 students (18–22 years old) with Turkish as their first language. Three samples were gathered over 48 weeks on three distinct themes from each group, rated across A2, B1, and B2 CEFR levels. Polat discovered significant results for clause length, sentence length, T-unit length, complex T-unit per T-unit, dependent clause per clause, T-unit per sentence, complex nominals per clause, complex nominals per T-unit, and clauses per sentence, showed that higher proficiency scripts had more significant mean values than the scripts rated at lower adjacent or non-adjacent proficiency levels, for example, CEFR levels B2 vs B1. In addition, clauses per T-unit and verb phrases per T-unit showed higher proficiency level scripts had higher mean values between B1 vs A1; and B1 vs A2. Coordinate phrases per clause grew significantly from A2 to B1 but decreased from B1 to B2.

Polat reported non-significant results for clauses per T-unit and verb phrases per T-unit for learners in A2 vs A1.

The study of SC and CEFR levels has been subject to limitations in previous research. Many studies have utilized a restricted number of SC indices and a limited corpus of writing samples. Moreover, most research has concentrated solely on learners from a single L1 background or a mixed background, prohibiting the comparison of learners from specific L1 backgrounds. Additionally, none of the previous studies have compared learners of different ages. While these studies have employed CEFR as a reference point, the accuracy of learner assignments and their writings on CEFR levels has varied and, at times, been of unknown or questionable quality.

To address these limitations, the current study has employed a comprehensive range of SC indices and a considerable corpus of writing performances. The study has also included EFL learners from two L1 backgrounds and various age groups. The researchers have taken significant care to ensure that the analyzed writing samples are situated on CEFR levels as reliably as possible.

Table 5.3 Studies which explored syntactic complexity with reference to the CEFR levels

Researchers	**Indices**	**The CEFR levels that the indices separated**
Kim. (2004)	Adverbial, adjective & nominal clauses per clause; clauses and dependent clauses per T-unit, dependent clauses per clause; Prepositional, participial, gerund and infinitive phrases per clause	A2 vs. B2 (more clearly between B1 & B2 than between A2 & B1)
Hawkins & Filipović (2012)	Mean sentence length	A2 vs. B1. B1 vs. B2 B2 vs. C1. C1 vs. C2
Green (2012)	Mean noun phrase incidence, the mean number of modifiers per noun; sentence syntax similarity	B2 vs. C1 C1 vs. C2
Qi (2014)	Mean sentence length, clauses per sentence and dependent clauses per sentence learners.	At B1.2 level EFL versus ESL
Verspoor et al., (2012)	Mean T-unit length	A1 vs. A2. A2 vs. B1
Gyllstad et al. (2014)	Mean T-unit length, mean clause length, clauses per T-unit	A2 vs. B1
Banerjee et al. (2015)	modifiers per noun phrase	C2 vs. below C2 level

(continues)

TABLE 5.3 continues

Researchers	**Indices**	**The CEFR levels that the indices separated**
Alexopoulou et al. (2017)	Sentence length; Mean length of the clause, subordinate clauses per T-unit	A1 / A2 to B2
Yoon (2017)	T-unit length, clause length and complex nominal per clause, T-unit length and clause length, complex nominal per clause	A2 vs. B2 B1.2 vs. B2 A2 vs. B2
Martinez (2018)	the mean length of sentences, noun phrase per clause, and complex sentence ratio. Compound sentence ratio, complex sentence ratio, compound-complex sentence ratio, coordinate clause ratio, dependent clause ratio. The mean sentence length, compound sentence ratio, complex sentence ratio, coordinate clause ratio, dependent clause ratio, and noun phrase per clause.	High correlations between global score and syntactic complexity measures. Moderate correlations between global score and syntactic complexity. Significant differences between 3rd year versus fourth-year students in syntactic complexity indices
Paquot (2019)	clauses per T-unit, dependent clauses per T-unit, dependent clauses per clause, mean length of the clause, verb phrases per T-unit, complex nominals per T-unit, and complex nominals per clause	B2 versus C1, C1 versus C2
Polat (2019)	the sentence, T-unit and clause length, complex T-unit per T-unit, dependent clause per clause, dependent clause per T-unit, T-unit per sentence, complex nominal per clause, complex nominal per T-unit, and clauses per sentence Clauses per T-unit and verb phrases per T-unit Coordinate phrases per clause Coordinate phrases per T-unit Clauses per T-unit and verb phrases per T-unit	B2 vs. B1, B1 vs. A2, A2 vs. B2. A2 vs. B1, B1 vs. B2 A2 vs. B1 B2 vs. A2 A2 vs. B1, B1 vs. B2 B2 vs. A2

6 METHODS

This chapter describes the research questions, data collection, tasks, and participants in the study. Furthermore, this chapter describes transcribing the scripts (corpus), the rating procedure, and correcting errors in the written scripts. Furthermore, it describes the data analysis tools and methodology, including linguistic and statistical analyses.

6.1 Research questions

The research aimed to shed light on how syntactic complexity relates to the CEFR levels in English as a foreign language learners' writing by investigating which features of SC differentiate between the CEFR levels. The research also investigated if such factors as learners' L1 and age were related to observed differences in SC between the CEFR levels.

The first study compared two groups of English language learners whose L1 and educational system differed but whose writing in English was at the same CEFR level. It had the following research questions:

- RQ1. What syntactic complexity features in argumentative essays by Sindhi and Finnish EFL learners distinguish between CEFR levels A1, A2 and B1?
- RQ2.Which syntactic complexity features differ or remain the same between the Sindhi and Finnish EFL learners when their CEFR writing levels are the same?

The second study was partly inspired by a study by Lu (2011), which investigated 14 syntactic complexity features in the writing of Chinese college-level learners of English. The study focused on Finnish learners and compared two age groups.

- RQ1. To what extent is the syntactic complexity in the writing of two age groups of Finnish EFL learners related to their EFL writing ability? Which SC indices correlate strongest with their ability, and do the two age groups differ?
- RQ2. Which SC indices distinguish Finnish EFL learners at different CEFR levels, and do the two age groups differ?

The third study explored the impact of the varied ages of EFL learners on syntactic complexity in the Pakistani and Finnish contexts. In Pakistan, at the A1, A2, and B1 proficiency levels, two age groups of learners with identical L1 and proficiency were compared. Two groups of learners of different ages with the same L1 and competency were compared in the context of Finnish at levels A2 and B1.

- RQ1. Do the Pakistani EFL learners of different ages but with the same CEFR proficiency level (e.g., A1, A2 or B1) differ in syntactic complexity? If yes, which aspects of syntactic complexity are affected?
- RQ2. Do Finnish EFL learners of different ages but with the same CEFR proficiency level (e.g., A1, A2 or B1) differ in syntactic complexity? If yes, which aspects of syntactic complexity are affected?

The focus in all three studies was on quantitative, numerical differences (and similarities) in syntactic complexity across the CEFR levels rather than on characterizing and illustrating typical linguistic features of each level (e.g., syntactic structures at each level), which is the focus of much of the research carried out in the English Profile project, for example.

6.2 Data collection from Finland and Pakistan

The dissertation is based on an analysis of English as a Foreign Language (EFL) writings authored by students in Finland and Pakistan. The Finnish EFL students were in the eighth and twelfth grades, whereas their Pakistani counterparts were in the eighth through twelfth grades. In Finland, the students completed three assignments, one shared by the age and grade groups and two unique to each group. The tasks required students to express an opinion and explain a rationale, while some required them to describe something. In Pakistan, students completed a more extensive and diverse set of six writing assignments, ranging from composing informal and semi-formal emails, letters, and messages to narrating and expressing viewpoints; however, only one of the latter tasks (one that required expressing an opinion and giving reasons for it) was included in

the current study. The reason for selecting only one of the tasks for the current study was to increase the comparability of the EFL writing analyzed in Studies 1 and 3 by including only one of the same writing tasks from both L1 groups. In Study 1, the comparison of the two L1 groups was an explicit focus of the study, and although Study 3 focused on comparing two age groups, this was done in parallel in the two L1 groups.

In both countries, the tasks were completed under the supervision of a researcher or a teacher, following precise instructions and time limits. The participants in both nations were given sufficient time to demonstrate their writing skills without undue intervention from factors such as weariness.

6.2.1 Tasks in Pakistan

This section provides descriptions of tasks written by EFL students from Pakistan. Before completing the tasks, the informants completed a background questionnaire providing information about their personal, contextual, and linguistic backgrounds. After that, each student completed six assignments for approximately seven thousand written scripts. The tasks came initially from two studies on the development of writing ability in Finland: the CEFLING study (2007-2009) and the TOPLING study (2010–2013; see https://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet/cefling and https://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet/topling) and were used either unaltered or with slight modifications (to make them suitable for the different cultural contexts) in Pakistan. For this doctoral dissertation, only one of the tasks was included (the same that was completed by both learner groups in the Finnish DIALUKI study from which the Finnish data for this dissertation came). This was the following task:

Write on the following topics: What do you think about the matter? Give reasons for your opinion.
Topic: No mobile phones at school! (Give at least a few reasons)
Write in English in readable characters in the space below (which continues the reverse side). Write at least 50 words.

As mentioned above, the Pakistani students also wrote five other tasks ranging from e-mails to narrative texts, including a version of the above task on mobile phones that required them to take a stand on whether boys and girls should go to different classes at school (see Task 3, option 1 in the description of the tasks that the Finnish EFL students completed, as shown in the following section).

6.2.2 Tasks in Finland

Each participant in Finland wrote three scripts characterized as descriptive, argumentative, and narrative (a total of approximately 1280 scripts). As described earlier, the 8th and 12th graders had two tasks that were unique to their group (tasks 1 and 2 below) and one that was the same for both groups (task 3 below). The learners also completed this last task in Pakistan. Tasks 1 and 2 in both groups came from the Pearson Test of English General. The DIALUKI project collaborated with Pearson (see Harding et al., 2015; Alderson, Huhta, & Nieminen, 2016).

Writing tasks completed by the 12th graders in Finland:

Task 1: Use the information in Task 4 to help you write your answer.
The magazine in Task 4, where you saw the article about rail and air travel, asked readers to give their opinions. Write a short letter in response to the article. Your letter should include
Your opinion about cheap air travel and the environment
Whether you prefer travelling by plane or train,
Give reasons for your opinions.
Write 70 to 90 words.

Task 2: You see this notice in an English-language magazine.
We are calling all travellers!
Wanted: Articles
We are looking for articles titled "A Journey I will never forget."
Include information on the following:
Where the journey was
What was so special about it?
Why do you think you will never forget it?
The best articles will be printed in next month's edition.
Write 100-150 words.

Task 3: (task instruction was originally in Finnish, but the titles of the two topics were in English)
An opinion
Please choose one of the two topics and share your thoughts. Give reasons for your opinion.
1. Boys and girls should go to different classes at school.
2. No mobile phones at school!
Write in English in the space below in clear characters. Write at least a few sentences.

The writing tasks completed by the eighth graders in Finland:

Task 1:
Use the information in Task 4 to help you write your answer.
You received James' email. Write an email in response. In your email, you should:
Tell James that he has your history textbook.
Tell him that you will bring sandwiches for lunch.
Tell him you still need to do the chemistry homework.
Ask him for help with this.
Write 50 to 70 words.

Task 2: *You see the following transport survey in your college magazine.*
How do you travel to college every day? Tick one and explain why.

Car □ Train □ Bus □ Bicycle □ Walk □ Other □

Your answer should include
· Why do you travel that way?
· What do you like and dislike about it?
· How you would prefer to travel to college
Write 100 to 150 words.

Task 3: (task instruction was originally in Finnish, but the titles of the two topics were in English)
An opinion
Please choose one of the two topics and share your thoughts. Give reasons for your opinion.
1. Boys and girls should go to different classes at school.
2. No mobile phones at school!
Write in English in the space below in clear characters. Write at least a few sentences.

6.3 Informants

The data in Pakistan were gathered from around 940 informants with Sindhi as their mother tongue. Sindhi EFL learners were from 8th to 12th grades and came from 31 government, semi-government, and private schools. The participants' range in age was from 13 to 18 years. Almost all the Sindhi EFL learners were in the 8th to 12th grades. There were 261 learners from the 12th; 228 from the 11th; 245 from the 10th; 210 from the 9th; and 267 from the 8th. The learners produced around seven thousand written scripts.

The Finnish data were collected from 208 eighth graders and 219 12th-graders who were learning English as a foreign language in Finnish-medium

schools in Finland. The former were lower-secondary students in the comprehensive school, and the latter were upper-secondary level students from the academically oriented branch of secondary education (from the so-called gymnasia). The eighth graders were 13 or 14, and the gymnasium students were 17 or 18. The participants had studied English since their third year in primary school. The participants came from 15 different lower secondary schools (the 8th graders) and nine different upper secondary schools (the 12th graders) who participated in the project DIALUKI at the University of Jyväskylä (see Alderson, Haapakangas, Huhta, Nieminen, and Ullakonoja, 2014, for details).

6.4 Transcribing the scripts (Corpus)

The handwritten tasks were transcribed into digital format and anonymised (original identity hidden) with a particular coded identity. Several corpora were created based on the learners' grade level: in Finland, there were two for the 8th graders and one for the 12th graders (these had been created during the DIALUKI project in 2010–13). In Pakistan, too, there were separate corpora for each grade level. The corpora were then divided into sub-sets, each consisting of scripts written for a particular task.

6.5 Rating procedure

All the tasks were coded and rated anonymously. The raters did not know the writers' gender, school, age, or level. The scripts' rating was carried out concerning the CEFR scale from A1 to C2 in both countries involved in this research (Pakistan and Finland).

The CEFR-based rating scale used in the study was a compilation of several writing scales included in the CEFR (see Huhta et al., 2014, for details; a copy of the rating scale used both in the DIALUKI study, from which the Finnish data come, and in data collection in Pakistan can be found in Appendix 2). The scales described different aspects of writing and writing tasks relevant to the study (both the Finnish DIALUKI study and the data collected in Pakistan). An essential point in selecting the scales to be used as the rating scale was to exclude such scales in the CEFR that focus on the linguistic aspects of learner performance, even in rather general and non-language-specific ways. This was done to minimise circularity in the ratings by making the raters focus on the communicative quality of the texts and not on the language used by them (see Khushik & Huhta, 2022, for a discussion of the issue of circularity in such studies as the current dissertation).

The Finnish performances were rated both on the six-point CEFR scale and the ten-point Finnish curriculum scale based on the CEFR scale. The proficiency levels of the scripts were established in Finland before the start of the current

study, as the Finnish data come from the DIALUKI study between 2010 and 2013 (see Alderson et al. 2015). The raters were experienced EFL teachers and teacher trainers, or researchers trained using rated benchmark samples from an earlier study on EFL writing in Finland (the CEFLING and TOPLING studies, referred to earlier in the section on the writing tasks). The scripts written by Finnish EFL learners were double- or triple-rated.

The rating procedure in Pakistan was based on the procedure developed in the above-mentioned Finnish project and applied to the Pakistani learners' written scripts. A total of twenty-four volunteers participated in the rating process in Pakistan. All the raters were currently teaching EFL in colleges and universities. To have a variety of raters, male and female raters were included. All the raters had a minimum of a master's degree in English as a qualification, and some were PhD holders. Initially, the Pakistani raters were trained on 30 samples collected in the Pakistani context. Ratings of the 30 samples by each rater were discussed in full sessions. Then the raters were given a selection of new written samples collected in Pakistan to rate with ample time. They were instructed to rate at least fifty written scripts daily to help them maintain consistent ratings. Each written task got a minimum of four ratings and a maximum of seven ratings from the raters. A selection of the Pakistani scripts was cross rated by two Finnish raters who had previously participated in rating the scripts of Finnish EFL learners to strengthen the comparability of CEFR-level ratings in the two countries.

The ratings were analysed statistically using the multifaceted Rasch analysis program Facets (Linacre, 2009). The Facets programme was explicitly designed to analyse subjective rating data to ensure dependable placement of performances such as written scripts on a scale (e.g., the CEFR levels).

Facets were used to check the raters' consistency and severity. Infit values (Engelhard, 1994) were used to control the quality of the ratings. As a result of the analyses, three misfitting extremely lenient/severe raters were excluded from ratings to increase the quality of the Pakistani rating data. The fair average values from Facets were then used to assign a CEFR level to each script by rounding the values to the nearest CEFR level (see Alderson et al. 2015; detailed Facets' analyses and rating procedures are also described in Khushik & Huhta 2020).

6.6 Correction of linguistic errors

It is the process of deriving linguistic traits from written scripts that constitutes linguistic analysis. By using this approach, one can identify and interpret the various elements of a language, such as syntax, morphology, and phonology. In order to gain a better understanding of how languages develop, linguists analyze language structures in order to uncover patterns that reveal how a language develops during different developmental stages. In addition, they can identify how language has developed over time. This can be used to better understand the linguistic characteristics of learners in any language and context.

Uncorrected learner-written corpora often contain common errors such as missing punctuation marks, spelling errors, and other errors that require clarification for automatic analysis. These errors can have a significant impact on the accuracy of the analysis results, as the software may not be able to distinguish between a correctly written sentence and an incorrectly written one. Additionally, the lack of punctuation can make it difficult for the software to accurately parse the sentence.

As a result, automated systems may incorrectly annotate the syntactic structures of the scripts due to their misleading syntactic structures. Thus, the linguistic analyses of the written scripts would not yield reliable results. Automated systems rely on algorithms to identify syntactic structures in written scripts. However, these algorithms may not accurately determine syntactic structures due to corpora errors. This can lead to incorrect annotations and inaccurate linguistic analysis results. This is akin to a machine trying to solve a mathematical equation without being able to identify the proper elements in the equation due to input errors. In both cases, the output of the system is highly dependent on the accuracy of the input data.

For this reason, corpus construction for linguistic analysis requires pruning. To ensure high quality output, it is essential to pay attention to the quality of the corpus used as input. Cleaning the corpus is necessary for reliable linguistic analysis. In order for automated systems to be able to accurately identify the language properties of learners' performance, corpora must be cleaned.

According to McNamara et al. (2014, p. 155-156), at least 95% of a corpus should be cleaned before linguistic features can be extracted. For this reason, it was necessary to carefully evaluate the scripts used in the present study in order to determine whether the automatic applications functioned as intended. For example, when automated systems construct the indices, fragments, typos, and missing punctuation will affect the results (see also Banerjee et al., 2007). The author of this dissertation examined sample texts drawn from actual data during this process. He examined the effects of including texts with and without deceptive typographical errors on syntactic complexity indices. There were numerous data errors that hindered automated programmes from accurately identifying the language characteristics of learners.

Therefore, punctuation marks were added to sentences if they were missing. The endnotes, topic words, and numbers (added by learners) were removed. The spelling errors were corrected. Additionally, students whose brief paragraphs were written in their native language were excluded from the study. This ensured that the sentences were all in the same language and that punctuation marks were standardized. Additionally, removing endnotes and topic words allowed sentences to be analyzed without bias. Lastly, correcting spelling errors helped to ensure that automated tools correctly understood the sentences. In addition, several students who plagiarized the instructions and topic of the assignment but did not produce original writing or wrote off-topic were removed from the analyses. By pre-processing the data in this way, it was easier to compare sentences across different assignments and to ensure that the results of

the analyses were not skewed by language or formatting differences or by plagiarism. Also, the author first analyzed a few scripts with and without errors. Second, the author of the dissertation has analysed some samples manually which are added with their results in the appendix 3. This allows the author to compare the script results with the manual analysis and identify any discrepancies. Manual analysis gives the author an opportunity to check the accuracy of the scripts, tools and make any necessary changes to ensure reliability.

7 ANALYSES

The dissertation included both linguistic and statistical analyses. Language analyses were used to extract the linguistic characteristics of each script the learners wrote. Statistical analyses revealed if there were significant differences in the syntactic complexity features across the CEFR levels or if the L1 or age groups differed from one another. The first section describes the linguistic analyses.

7.1 Natural language processing and learner corpus research

The analysis of human language with natural language processing (NLP) programmes that use machine learning approaches has been revolutionized (see Kyle, 2021, and McNamara, 2021). The use of NLP methods to examine complicated linguistic phenomena in learner corpora of second and foreign-language learners is increasing. Complex linguistic events and linguistic properties (for example, vocabulary, grammar, and coherence) are becoming increasingly important for understanding, teaching, and evaluating second and foreign-language learners. The analysis of human language with natural language processing (NLP) programmes that use machine learning approaches has been revolutionized (see Kyle, 2021, and McNamara, 2021). The use of NLP methods to examine complicated linguistic phenomena in learner corpora of second and foreign-language learners is increasing. Complex linguistic events and linguistic properties (for example, vocabulary, grammar, and coherence) are becoming increasingly important for understanding, teaching, and evaluating second and foreign-language learners. NLP methods are able to provide quantitative measures of these important properties that can then be used to identify patterns and trends in learners' language, which can help inform and improve language teaching and evaluation. By leveraging NLP methods, researchers can uncover and analyze patterns in language learners' use of

vocabulary, grammar, coherence, and other complex linguistic events, which can provide valuable insights for advancing language teaching and evaluation.

Kyle (2021) provides a comprehensive overview of the NLP and learner corpus research. According to Kyle (2021, p. 1), "natural language processing" is a method for mechanically analyzing human language. Kyle (2021) explains that NLP involves using computers to recognize patterns in language, convert text into structured data, and generate language-based solutions. He also discusses the importance of learner corpus research, which is the study of language usage in various contexts, and how it can inform the development of NLP applications.

Kyle (2016) further explains that the processes that fall under the NLP umbrella can range from those that are relatively straightforward, such as separating character sequences into words and sentences, to those that are significantly more complex and challenging, such as converting speech into text and annotating texts for syntactic, semantic, and pragmatic features. Simple tasks such as turning letter sequences into their component words and sentences can be performed relatively quickly and accurately. However, the more complex analyses can be challenging. NLP methods of varying degrees of complexity have substantially increased corpus linguistics and learner corpus research. In addition, they have increased the use of corpora to investigate language learning. These methods involve Artificial Intelligence algorithms that can analyze large amounts of data to identify patterns and relationships between words and phrases. These algorithms can be used to identify the most common words, phrases, and topics in a corpus, to uncover similarities and differences between two different corpora, and to identify how language use changes over time. Furthermore, such algorithms can be used to analyze learners' language use in order to identify areas of strength and areas of improvement, and to assess the effectiveness of language learning strategies.

7.2 Linguistic Analysis to extract linguistic features

Linguistic analysis is the examination of linguistic characteristics contained within language data. Language can be analyzed from a variety of angles. For instance, a piece of writing can be analyzed to investigate its semantics, coherence, syntax, and lexis. Historically, linguistic analysis was carried out manually using various forms and techniques. Language properties, structures, and forms are now computed electronically by computer applications, which saves a considerable amount of time and energy.

The current dissertation used two up-to-date natural language processing tools. The first was the L2 Syntactic and Lexical Complexity Analyser (http://www.personal.psu.edu/xxl13/download.html), which calculates 48 syntactic and lexical complexity features. The second was the Coh-Metrix (http://cohmetrix.memphis.edu), which calculates 106 linguistic features (related to, for example, cohesion, semantics, lexical diversity, connectives, syntactic complexity, syntactic pattern density, word information, and

readability) of the texts. For the dissertation, only those indices related to syntactic complexity were used. The reasons for using these two tools relate to their wide application in previous research on syntactic complexity and linguistic features more generally and to the fact that the two, combined, covered a wide range of SC features investigated in previous studies on SC, some of which were less widely used syntactic features.

The explored syntactic complexity features represented three levels (i.e., sentence, clausal, and phrasal levels) in figures 7.1, 7.2, and 7.3 below. Additionally, Table 4 of Appendix 1 further clarifies what is meant in this dissertation when reference is made to various indices as constructs (e.g., clauses, T-units).

Figure 7.1. Syntactic complexity measured at the sentence level in the current dissertation

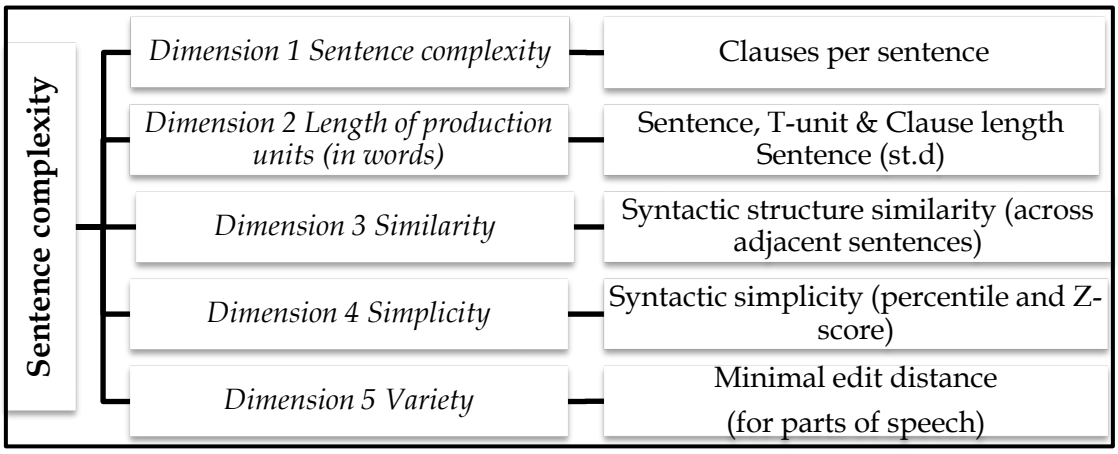


Figure 7.2 Syntactic complexity measured at the clausal level in the current dissertation

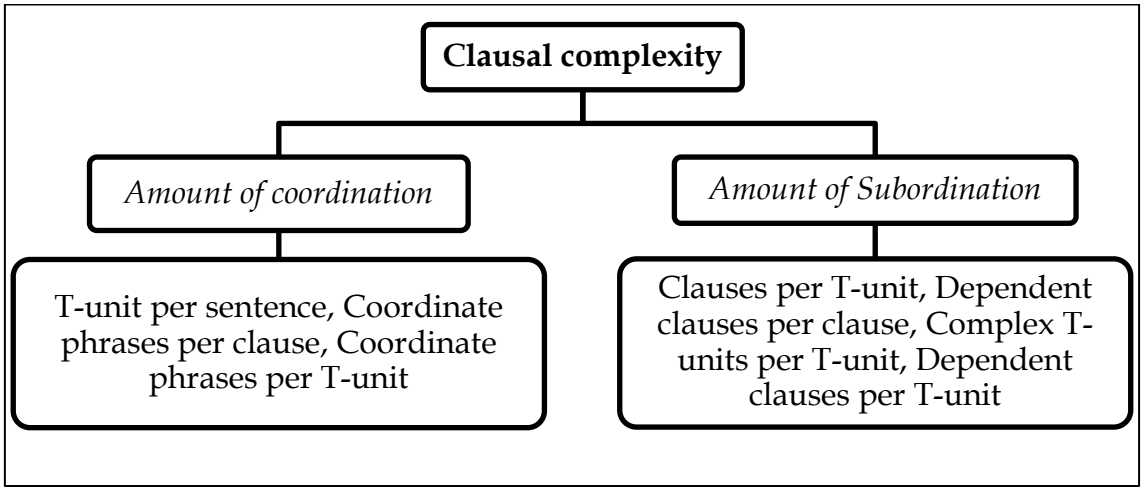
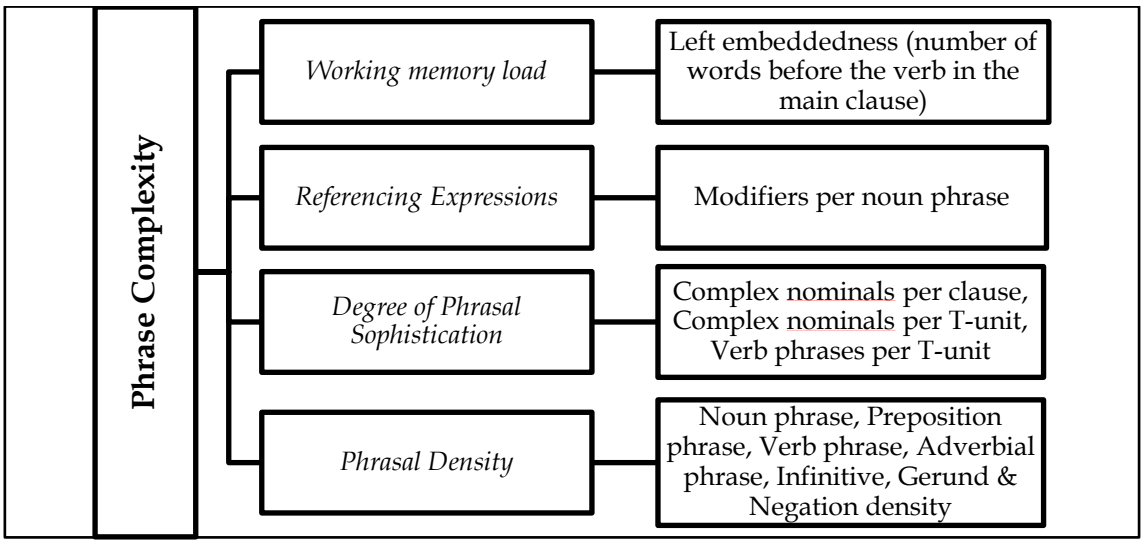


Figure 7.3 Syntactic complexity measured at the phrasal level in the current dissertation



7.3 Definitions of the SC measures used in this study

SC measures fall into two categories: (1) clauses, sentences, and T-units and their relationships with one another; and (2) specific grammatical structures (e.g., passives, nominals) relating to clauses, sentences, or T-units (see, e.g., Wolfe-Quintero et al., 1998). The four basic approaches for calculating syntactic complexity features are frequencies, ratios, percentages, and indices. The frequency of a grammatical structure is the count of its occurrences. The ratio is calculated by dividing the number of one type of unit by the total number of another type of unit. A percentage (%) is a relative measure expressing hundredths of a quantity. An index is calculated using algorithms to arrive at numerical values (Wolfe-Quintero et al., 1998), for example, the number of occurrences of a particular structure per 1000 words.

Next, the most commonly used syntactic units (e.g., sentence, clause, T-unit) are defined (Table 7.3). This is followed by describing and defining the syntactic complexity indices investigated in the current study. Since two tools, the L2 Syntactic Complexity Analyzer and Coh-Metrix, were used in the study, and the SC indices are presented separately in two tables, one for each of the tools (Tables 7.4 and 7.5). See also Table 4 of Appendix 1 on the definition of the various indices as constructs.

Table 7.3. Definitions of the key syntactic units of interest in studies of syntactic complexity

Syntactic units / concepts	**Definition of the syntactic unit**
Sentence	Group of words punctuated by the writer (Hunt, 1965; Tapia, 1993). A sentence is defined as a group of words (including sentence fragments) punctuated with a sentence-final punctuation mark, including a period, exclamation mark, question mark, and occasionally elliptical marks or closing quotation marks.
T-unit	T-unit is defined as "one main clause plus the subordinate clauses attached to or embedded within it" (Hunt, 1965, p. 49)
Clause	Clause: a phrase dominated by VP or S (Bardovi-Harlig & Bofman, 1989) OR a structure with a subject and a finite verb (Hunt, 1965; Polio, 1997). A clause is a structure with a subject and a finite verb, including independent, adjective, adverbial, and nominal clauses, but not non-finite verb phrases, which are included in the definition of verb phrases instead (Hunt 1965; Polio 1997) (LaBrant 1931 first introduced the clause as the unit of study).
Dependent clause	A dependent clause is a finite adverbial, adjective, or nominal clause (Cooper, 1976; Hunt, 1965; Kameen, 1979).
Complex T-unit	A complex T-unit is one that contains a dependent clause.
Coordinate phrases	Phrases composed of two or more elements (such as noun phrases or verb phrases) that are connected by a coordinating conjunction, e.g., She saw a red car and some birds (here 'a red car' is a phrase and 'some birds' is also a phrase). When we insert a coordinating conjunction such as "and" between them, we have (apparently) two coordinate sentences. Lu (2010) counts only adjectives, adverbs, nouns, and verb phrases in coordinate phrases.
Complex nominals	"Complex nominals comprise (i) nouns plus an adjective, possessive, prepositional phrase, relative clause, participle, or appositive, (ii) nominal clauses, and (iii) gerunds and infinitives in subject position " (Lu 2010, p. 483)
Verb phrases	"Verb phrases comprise both finite and non-finite verb phrases e.g., is 'acting like a spoiled child' where 'acting like' and 'a spoiled child' are verbs" (Lu 2010, p. 484)

Coh-Metrix covers various linguistic and textual features, including several indices relevant to syntactic complexity (Graesser et al., 2004). These include the mean length and standard deviation for such production units as words, sentences, and paragraphs. In addition, Coh-Metrix focuses on syntactic (dis)similarity by measuring uniformity and consistency, sentence-to-sentence syntax similarity, and syntactic pattern density. Indices analyzed through *Coh-Metrix* are listed in Table 7.4. Coh-Metrix is an online tool, and the analyzed texts were input into the available window on the Coh-Metrix website.

Table 7.4 Definitions of the syntactic complexity indices in Coh-Metrix (Graesser et al., 2004) (Source: http://www.cohmetrix.com/)

Index	**Definitions of the indices from Coh-Metrix (Graesser et al., 2004), McNamara et al., (2014)**
The standard deviation of the mean length of sentences	"This is the standard deviation of the measure for the mean length of *sentences* within the text. A large standard deviation indicates that the text has a large variation in terms of the lengths of its sentences, such that it may have some very short and some very long sentences. The presence of headers in a short text may impact this measure. Narrative text may also have variations in sentence length as authors move from short character utterances to long descriptions of scenes" (Graesser et al., 2004)
Syntactic simplicity (z-score) (percentile)	"This component reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures that are less challenging to process. For example, some sentences are short and have a simple syntax that follows an actor-action-object syntactic pattern, have few if any embedded clauses, and have an active rather than passive voice. The syntax in text tends to be easier to process when there are shorter sentences, few words before the main verb of the main clause, and few words per noun-phase. At the opposite end of the continuum are texts that contain sentences with more words and that use complex, unfamiliar syntactic structures…" (Graesser et al., 2004)
Left embeddedness	The mean number of words before the main verb. "The left embeddedness (SYNLE) is embedded constituents with difficult syntactic constructions and is often structurally dense, syntactically ambiguous, or ungrammatical used in sentences …. As a consequence, sentences are more difficult to process and comprehend …." (Graesser et al., 2004) (Perfetti, Landi, & Oakhill, 2005).
Modifiers per noun phrase	"A word, especially an adjective or noun used attributively, that restricts or adds to the sense of a head noun (e.g., *good* and *dog* in *a good doghouse*) …" (Graesser et al., 2004)

continues

Table 7.4 continues

Index	**Definitions of the indices from Coh-Metrix (Graesser et al., 2004), McNamara et al., (2014)**
Minimal edit distance for parts of speech	"Coh-Metrix provides Minimal edit distance for parts of speech. MED calculates the average minimal edit or the distance that parts of speech are from one another between consecutive sentences in a text…" (Graesser et al., 2004). "Example: the dog chases the cat. The cat chases the dog. The syntactic dissimilarity is 0.0 because the syntax is the same, considers parts of speech but not the words themselves (e.g., determiner + noun) …" (Graesser et al., 2004)
Sentence syntax similarity (adjacent sentences & across paragraphs)	"Sentence syntax similarity (adjacent sentences) is the average parse tree similarity between all combinations of sentence pairs across paragraphs of the text. Sentence syntax similarity (across paragraphs) is based on parse tree similarities between sentences. For two-sentence parse trees, the maximum common tree is found by removing uncommon subtrees… " (Graesser et al., 2004).
Syntactic Pattern Density indices	"Syntactic complexity is also informed by the density of particular syntactic patterns, word types, and phrase types. Coh-Metrix provides information on the incidence of noun phrases (DRNP), verb phrases (DRVP), adverbial phrases (DRAP), and prepositions (DRPP). The relative density of each of these can be expected to affect processing difficulty of text, particularly concerning other features in a text. For example, if a text has a higher noun and verb phrase incidence, it is more likely to be informationally dense with a complex syntax, Negation (DRNEG) is also associated with processing difficulty …" (Graesser et al., 2004) Coh-Metrix provides an indicator regarding the incidence of verb conjugation in the text. It provides the relative frequency of the use of the gerund (DRGERUND; in its -ing form) as well as verbs as infinitives (DRINF). A verb's infinitive is its unmarked form, such as "be," "have," or "write." Infinitives are prevalent in situation models with a high density of intentional content, where agents perform actions in order to achieve goals. …" (Graesser et al., 2004)

The L2 SCA calculates 25 indices covering syntax and lexis, 14 of which focus on syntax and are based on the second-language development literature. Both online and offline versions of L2SCA exist, and the current study used the offline version since it was easier to input many texts into it. The analyzer accepts a plain text sample of written English and generates the 14 indices of syntactic complexity for the sample. Lu (2010) argues that the systems that preceded L2SCA were insufficiently valid and reliable for increasingly complex, profound, protracted, and syntactically advanced utterances. Thus, the length of the production, general complexity measurements, the level of coordination, syntactic dependency, and the presence of specific grammatical structures are all factors considered by L2SCA.

Table 7.5. Syntactic complexity indices calculated by L2 Syntactic Complexity Analyzer

Index	**Calculation method or formula (Lu 2010, 2011, Ai & Lu, 2013)**
mean length of sentence	number of words / number of sentences
mean length of t-unit	number of words / numbers of T-units.
mean length of clause	number of words / number of clauses
clauses per sentence	number of clauses / number of sentences
t-unit complexity ratio	number of clauses / numbers of T-units
complex t-unit ratio	number of complex t-units / number of T-units
dependent clause ratio	number of dependent clauses / number of clauses
dependent clauses per T-unit	number of dependent clauses / numbers of T-units
coordinate phrases per clause	number of coordinate phrases / number of clauses
coordinate phrases per T-unit	number of coordinate phrases / number of T-units
sentence coordination ratio	number of T-units / number of sentences
complex nominals per clause	number of complex nominals / number of clauses
complex nominals per T-unit	number of complex nominals / numbers of T-units
verb phrases per t-unit	number of verb phrases / number of T-units

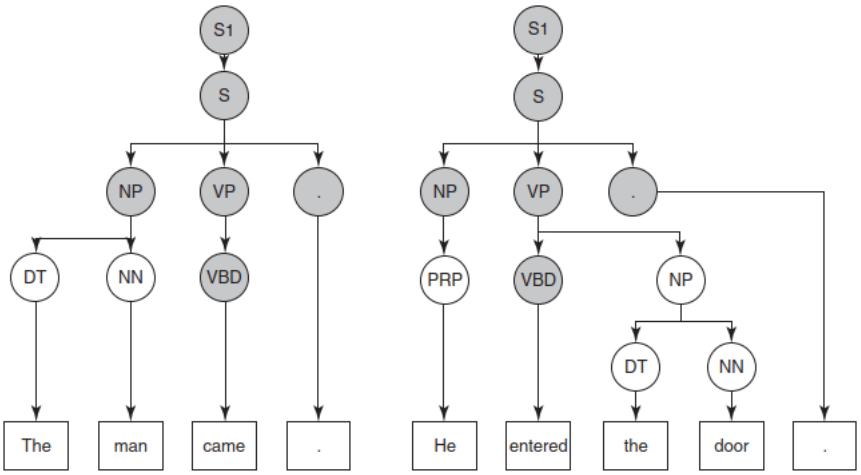
L2SCA and Coh-Metrix produced data, that is, values for each SC index and each text in .csv format. The data were then taken to SPSS for statistical analyses.

At this point, it is useful to elaborate on three SC indices that Coh-Metrix calculates, as they are less transparent than the other indices: *syntactic simplicity, sentence syntax similarity, and minimal edit distance (parts of speech).* McNamara et al. (2014, p. 86) define *syntactic simplicity* as the degree to which sentences in a text contain fewer words and use simpler, more familiar syntactic structures that are easier to understand. They reported that as proficiency increases, this index decreases (McNamara et al., 2014, pp. 254, 259, & 264).

According to McNamara et al. (2014, p. 71), Coh-Metrix produces metrics of *sentence-to-sentence syntax similarity* by measuring the regularity and consistency of syntactic structures found across the text. The sentence-to-sentence syntactic similarity is the average parse tree similarity (Sim) of adjacent sentence pairs inside a text. The maximum common tree can be found by starting with two different sentence parse trees and then removing any subtrees that are not typical of either tree. In order to determine the degree of similarity between two parse trees, the formula that follows is utilised:

Sim = nodes in a shared tree / (the sum of the nodes in the two sentence trees – nodes in a common tree)

Figure 7.4 Illustration of syntactic similarity (McNamara et al, 2014, p.71)



Note: DT = determiner, NN = noun (singular or mass), NP = noun phrase, PRP = personal pronoun, S1 = sentence, S = simple declarative clause, VBD = verb (past tense), VP = verb phrase

According to McNamara et al. (2014, p. 71), this figure illustrates the sentence-to-sentence syntax similarity that exists between two sentences that are adjacent to one another: "The man came. He entered the door" (McNamara et al., 2014, p. 71). The grey nodes in the diagram represent the nodes shared by both sentences. According to this example, six nodes have a common edge, while the remaining twelve do not, yielding an index value of 0.50.

The illustration shows how the structure of a typical tree looks. The initial tree has a total of 8 nodes, while the subsequent tree has a total of 10 nodes. In the diagram, the nodes that are considered to be common are coloured. There are six nodes that are shared by the two connections. Word-containing rectangle leaves do not contribute to the total count of nodes in the tree. As a direct consequence of this, the formula Sim = 6/ ((8+10)-6)) = 6/12 = 0.50 is used in the process of computing the similarity. This index considers the syntactic similarity of phrase pairs at the phrasal level and the components of speech involved. When syntactic constructions are more uniform, the resulting syntax is simpler, making it simpler for the reader to understand (Crossley, Greenfield, & McNamara, 2008).

The third index is Minimal Edit Distance for *Parts of Speech* (McNamara et al., 2014, p. 70). By measuring the uniformity and consistency of the sentence constructions in the text, this index (see McCarthy, Guess, & McNamara, 2009), Coh-Metrix evaluates a combination of semantic and syntactic dissimilarity in the text. This is accomplished by comparing the sentences in the text. It does this by computing the distance that separates successive sentences in a given text in terms of parts of speech and words. McNamara et al. (2014, p. 70) illustrate the index with this pair of sentences:

The dog chases the cat.
The cat chases the dog.

Since there is no difference in syntax, the minimal edit distance for parts of speech syntactic dissimilarity is set to 0.0, which focuses on the parts of speech rather than the words themselves (for example, the determiner and the noun). In its most basic form, the Minimal Edit Distance for Parts of Speech algorithm determines the degree to which a given sentence needs to be altered (edited) in order for it to have the same syntactic composition as another sentence.

In addition, it is important to remember that it is not common for Applied Linguistics studies to use many indices. This is due to the fact that most Applied Linguistics studies are qualitative in nature. There are very few researchers who are interested in combining multiple areas and methods. Moreover, these indices have been introduced recently, are more complex, and require further research and understanding. Furthermore, these indices are explored through tools which are based on alogrithims programmed via machine learning techniques, parsing, and annotation, and many researchers do not have the time or resources to gain such a thorough understanding of each index. In addition, the indices may not be applicable to the specific research questions being addressed. Furthermore, the data being collected may not be suitable for the application of indices. Moreover, the data collected may not be relevant to the research being conducted or outdated and therefore not as reliable. Additionally, the complexity of the data and the number of variables involved can make it difficult for researchers to identify meaningful patterns. In these cases, it may be more appropriate to use a limited number of indices.

The present dissertation utilizes the Holistic Syntactic Complexity Indices, a set of metrics that enables a comprehensive evaluation of the complexity of written texts. These indices represent a valuable resource for gauging the level of difficulty of a text and can be applied to a wide range of genres and lengths. Their versatility allows for pinpointing areas in need of improvement, assessing text proficiency levels, and comparing complexity measures across different languages. The evaluation system based on holistic indices is known for its reliability and efficiency, facilitating cross-level and cross-context comparisons. Holistic indices are indispensable tools for language learners striving to enhance their writing skills, as well as for educators and linguists researching text complexity. Empirical evidence has demonstrated that these indices represent reliable and valid measures of text complexity, highlighting their relevance for language development and instruction.

This dissertation aims to provide a more comprehensive perspective on syntactic complexity (SC) indices, given the study's timing in the mid-2010s when automated analysis tools like L2SCA and Coh-Metrix were most widely used. The author of this dissertation found Coh-Metrix's SC-related indices to be particularly useful in expanding the range of SC indices used in the study. Notably, these indices included phrasal-level indices that were not part of L2SCA.

This dissertation builds upon earlier research that used broader, more holistic SC indices, such as those reviewed by Ortega (2003) and Wolfe-Quintero (1999) in L2SCA.

In order to gain a better understanding of the differences in syntactic complexity (SC) across Common European Framework of Reference (CEFR) levels, we plan to utilize Kyle's innovative tools for data analysis in the future. This topic is further discussed in the final section of the dissertation. Our next research phase will involve implementing recommended indices by Kyle & Crossley (2018) and Biber et al. (2014) to scrutinize specific aspects of syntactic complexity in relation to CEFR levels. We will take into account various factors such as learner, task, and context to establish a more comprehensive comprehension of the syntax used by English as a Foreign Language (EFL) learners at different CEFR levels. The usage of current indices will allow for more accurate data analysis and identification of significant patterns. Therefore, investigating syntactic complexity using global and general measures based on CEFR levels is essential at this stage.

7.4 Multidimensional analysis

More recently, syntactic complexity is considered a multidimensional construct. The concept of syntactic complexity as a multidimensional construct in relation to CEF levels in written texts of EFL learners has rarely been examined except by Khushik and Huhta (2020 & 2022). This is like looking at a multi-faceted diamond. Each facet reflects a different aspect and, when taken together, a single image emerges that is far more meaningful than any single facet. Hence, the current dissertation used a multidimensional approach to studying the syntax of EFL students. Bulte and Housen (2012); Lu (2011); Lu and Ai (2015) and Norris and Ortega (2009) recommend incorporating all possible types, dimensions, and kinds of syntactic complexity indexes, as including a variety of indices can provide a more realistic picture of how syntactic complexity contributes to language development. By including a variety of indices, researchers can gain a better understanding of what factors play a role in the development of language. In the case of this study, the syntax of EFL learners was examined. Multidimensional analysis allows researchers to look at multiple factors simultaneously, providing a more comprehensive understanding of the data. Furthermore, this approach can also provide more detailed insights into the development of learners' syntactic abilities over time. Syntactic complexity refers to the number of syntactic units in a sentence (i.e., length), the number of different types of syntactic units (i.e., complexity of structures), and the use of different language features. By quantifying the syntactic complexity of written texts, we can understand how language development progresses in individuals. This is similar to a scaffolding process, where each step-in development provides a foundation and support for the next step in the process. This gradual progression helps the learner gain a more comprehensive understanding of the material over time.

Initial research on SC employing multidimensional analysis identified various dimensions in written texts. Carroll (1960) investigated written English

prose and discovered six dimensions, essentially statistically analyzed characteristics. However, Marckworth and Baker (1974) discovered three dimensions. Based on these earlier studies, Biber began conducting what he called a macroanalysis of texts in the mid-1980s to find out the dimensions of the text, which are the basic concept of macroscopic analysis and constitute the vital theoretical constructions (Biber, 1985, p. 338). Such macroscopic investigations shed light on the overall dimensions of linguistic diversity within a specific domain, such as printed English texts (Biber, 1985, p. 338). In the latter stages of his investigation, Biber utilized factor analysis (Biber, 1985, pp. 86–88) and introduced the concept of multidimensional analysis (Biber, 1988).

7.5 Statistical analyses

The statistical analysis of the data was carried out using SPSS. MANOVA, correlations, T-tests and descriptive statistics were used to address the research questions in each sub-study, that is, to examine if syntactic complexity was related to the CEFR levels A1, A2, B1, and B2 and if the levels differed in terms of SC in the learner groups involved in the study.

As part of the data cleaning process for each index, SPSS data normality tests were undertaken first, removing two to five univariate outliers per SC index or variable. The data were additionally examined for multivariate outliers using Mahalanobis distance analysis, eliminating a limited number of additional learners. Following this, a series of multivariate analyses of variance were performed for each dimension or combination of indices (e.g., count variables or variables related to phrasal level indices) to determine overall differences between CEFR levels. After this, univariate analyses and pairwise comparisons were conducted to determine which indices distinguish the CEFR levels. Similarly, Pearson correlation coefficients were utilized to evaluate the association between SC indices and judgments of writing skills (i.e., learner ability measures from Facets).

8 SUMMARY OF THE RESULTS

The dissertation's principal purpose was to examine the syntactic differences between and within the CEFR levels A1, A2, B1, and B2 in English as a foreign language. The evaluation of the texts using around 30 syntactic complexity indices provides us with objective, empirical information about linguistic comparability and progression across a wide variety of SC indices and several CEFR skill levels.

This section presents the overall summary of the results. The three studies (included at the end of this document) explored syntactic complexity concerning the CEFR levels, particularly differences in SC between the levels. The Pakistani and Finnish EFL learners involved in the study wrote several texts of different types and on different topics. However, the three studies that comprise this dissertation each used somewhat different combinations of the texts as data, depending on the research questions addressed in the particular study. The findings showed significant changes in most syntactic complexity features across the CEFR levels. The findings indicate, for example, that Pakistani learners' results showed significantly different syntactic complexity features across the CEFR levels compared with Finnish EFL learners.

Table 8.1 Key aspects of the three studies

Research questions of the study	L1 group / groups studied in the particular article as well as their size	Information about the task or tasks included in the study
Study number 1		
RQ1. What syntactic complexity features in argumentative essays written by Sindhi and Finnish EFL learners distinguish between CEFR levels A1, A2 and B1? RQ2. Which syntactic complexity features differ or remain the same between the Sindhi and Finnish EFL learners when their CEFR writing levels are the same?	Finnish L1: 287 8th and 12th graders. Sindhi L1: 868 8th to 12th graders.	One task on the same topic (argumentative essay) rated at A1, A2, and B1 levels
Study number 2		
RQ1. To what extent is the syntactic complexity in the writing of two age groups of Finnish EFL learners related to their EFL writing ability? Which SC indices correlate strongest with their ability, and do the two age groups differ? RQ2. Which SC indices distinguish Finnish EFL learners at different CEFR levels, and do the two age groups differ?	Finnish L1: 202 8th graders and 195 12th graders.	Three tasks (argumentative, persuasive/descriptive and narrative; but merged into one writing ability measure for each student) rated at A1, A2, B1, and B2 levels
Study number 3		
RQ1. To what extent does the syntactic complexity of Pakistani EFL learners of different ages with the same CEFR competency level (A1, A2, or B1) differ? RQ2. To what extent does the syntactic complexity of Finnish EFL learners of different ages with the same CEFR competency level (A1, A2, or B1) differ?	Finnish L1: 148 8th graders and 136 12th graders. Sindhi L1: 296 8th graders and 415 12th graders.	One task on the same topic (argumentative essay) rated across A1, A2, and B1 levels

8.1 What picture does each of the three studies paint?

This dissertation comprises three studies that offer a comprehensive understanding of the relationship between CEFR levels in EFL writing and syntactic complexity. The first study examined the impact of different syntactic

complexity features on CEFR levels and explored the effect of L1 (Sindhi or Finnish) on these features. This study revealed significant differences in SC features between the two L1 groups, which distinguished the writing at different CEFR levels. The second and third studies complemented the first one by analyzing the correlation between SC and writing ability in two age groups in one L1 group (Finnish). Also, the second study investigated differences in SC between the CEFR levels in two age groups in one L1 group. Unlike the first study, the second study used three writing tasks written by the learners to place them at CEFR levels. The third study focused on age-related differences in SC in both L1 groups and aimed to determine if some SC indices differed based on learners' age, even if their CEFR level did not.

8.1.1 First study

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across the Common European Framework of Reference Levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532.

The first study used a data set with essays on one topic (argumentative) rated at the CEFR proficiency levels A1, A2, and B1. The data from Pakistan were collected specifically for this dissertation. In contrast, the Finnish data had been gathered earlier in the Dialuki study at the University of Jyväskylä in 2010–2013 (see Sections 6.2.1 and 6.2.2). The study explored around 30 syntactic complexity measures to study the SC as a multidimensional construct. The study investigated whether the learners' syntactic complexity features differed across A1, A2, and B1 levels. The second aim was to explore whether SC differed in the two L1 groups (Finnish versus Sindhi) when their overall writing level was the same. This highlighted the CEFR levels' linguistic similarity across different L1 groups. Informants were adolescents from Pakistan (N = 868) and Finland (N = 287) who wrote the same argumentative essay and were graded on the CEFR-based scale A1-C2. Multifaceted Rasch analysis was used to analyse the ratings, and the L2 Syntactic Complexity Analyzer and Coh-Metrix were used to analyse the essays for SC indices.

Several SC indices distinguished the CEFR levels A1, A2, and B1 in both language groups. However, the length of production units, sentence complexity, subordination, and density indices produced the most pronounced distinctions between the CEFR levels. The most significant differences between the CEFR levels were the length of the measures and phrasal density. In the Sindhi group, the length of sentences and T-units, the number of verb phrases per T-unit, and the density of negative expressions were the clearest separators of CEFR levels. The most significant variations across CEFR levels in the Finnish group were identified for the mean sentence length, the standard deviation of the mean sentence length, the mean T-unit length, the number of verb phrases per T-unit, and the density of noun phrases.

Study 1 also demonstrated how each CEFR level could be characterised concerning the numerical values of the different SC indices (e.g., the typical mean length of sentence at each level) across EFL learners from various L1 backgrounds by using selected SC indices as examples from the first study and the study by Banerjee et al. (2007). This approach would complement the one employed in Study 1, which examined differences in SC between CEFR levels.

Regarding differences in SC between the two L1 groups of EFL learners (Sindhis vs. Finns), it was found that around half of the syntactic complexity indices out of twenty-nine showed significant differences between Sindhi and Finnish EFL learners at the A1 level. Similarly, more than twenty syntactic complexity differences were found between the L1 groups at the A2 and B1 levels. Thus, many indices distinguished the two language groups at the CEFR levels A1, A2, and B1. However, the clearest separators (regarding effect sizes) were the length of production units, subordination, and phrasal density indices. Several indices, however, remained constant, and the A1 level was more similar in both language groups compared with the A2 and B2 levels.

8.1.2 Second study

Khushik, G. A., & Huhta, A. (2022). Syntactic complexity in Finnish-background EFL learners' writing at CEFR levels A1–B2. *European Journal of Applied Linguistics*, *10*(1), 142-184.

The second study focused only on the Finnish data set. This data set contains the data of 12th-grade learners and 8th-grade learners collected in the DIALUKI research project in 2010-13. This study aimed to see if there was a link between the syntactic complexity of Finnish EFL learners' writing and their ability to write in EFL. Furthermore, the study aimed to discover which SC indices showed the strongest correlation with their writing ability and whether the two age groups differed. Second, the aim was to know which SC indices best differentiated Finnish EFL learners at different CEFR levels and whether some of the indices differed across the two age groups.

The second study focused on the linguistic profiles of 208 8th graders (aged 14) and 219 upper-secondary school students (aged 17) who completed three English writing tasks (narrative, argumentative, descriptive) from an international English language examination and a previous research project (see Section 6.2.2 for details). The writing tasks were assessed according to the CEFR scale A1-C2. The ratings were analysed using multifaceted Rasch analysis, and the texts were analysed using the same automated methods as in Study 1.

The first research question investigated the correlations between syntactic complexity indices and the learners' overall writing ability based on their collective performance across all three tasks. Significant and robust positive correlations were found between twenty indices and the learners' writing ability. The number of words, clauses, T-units, and sentences were all related to the length of the text, which has been linked to judgments of L2 writing quality in

many previous studies. The number of words, the most comprehensive metric of text length, corresponded most strongly with writing competence in both groups (grade 8 & gymnasium). However, in both groups, counts of all other linguistic units also corresponded significantly with ability. The number of complex nominals and phrases was also a good predictor. There were also variances between the groups, with the 12th graders writing more sentences and dependent clauses. The most noticeable difference, however, was that correlations between writing ability estimates and all count variables were substantially more robust in grade 8.

The second research question investigated whether certain syntactic complexity features separated different CEFR levels in age and grade groups. All count indexes significantly separated the CEFR levels. Separation was more visible in grade 8, as seen by more significant effect sizes than in the 12th graders in the gymnasium. The number of words students wrote increased significantly across practically all adjacent CEFR levels. In grade 8, almost all count variables distinguished between A1 and A2 learners on the one hand and A2 and B1 writers on the other.

In contrast, these factors, particularly the number of complex nominals, complex T-units, phrases, and sentences, did not separate A2 from B1 in the gymnasium but distinguished B1s from B2s. The number of words, clauses, sentences, and phrases increased continuously across levels, as suggested by the relatively high correlations between count variables and writing competence. The findings indicate that in the lower CEFR levels, the mean sentence and T-unit lengths, variance in sentence length, infinitive density, clauses per sentence or T-unit, and verb phrases per T-unit were the most apparent separators (A1 vs. A2). At higher levels (B1 vs. B2), modifiers per noun phrase, mean clause length, complex nominals per clause, and left embeddedness were clear separators between the levels. The results corroborate prior findings that the length and variance of more extensive production units (sentences, T-units) are the SC indices that most clearly differentiate the lower CEFR levels. In contrast, the higher levels are best differentiated by clausal and phrasal complexity.

8.1.3 Third study

(*Syntactic complexity in young and adult EFL learners' writing - Does the age matter when the learners' CEFR level is the same*? Khushik, Ghulam Abbas) '**submitted**'.

The third study investigated whether syntactic complexity differences exist between two age groups whose CEFR proficiency level is identical. The data set came from Pakistan and Finland (Sections 6.2.1 and 6.2.2). None of the previous studies has focused on learners with the same (CEFR) proficiency levels, which, however, differ in their age, and attempted to answer whether, for example, A1 level learners from the 8th grade are the same in terms of syntactic complexity as A1 level learners from the 12th grade. Like learners' L1 (see Study 1 above),

learners' age might affect SC in their EFL writing even if their general, communicatively oriented CEFR levels are the same.

This study aimed to determine whether the linguistic characteristics of young English language learners' writing performances were similar when the performances were graded at the same competence level. However, the learners' age (and grades) was different. The study examined the syntactic complexity of learners from Pakistan and Finland who represented two age groups and compared them at the CEFR levels A1, A2, and B1. Over a thousand EFL students in the eighth and twelfth grades wrote an English essay on an argumentative topic. The same two automated tools used in Studies 1 and 2 were used to analyse the writings for multidimensional syntactic complexity. Syntactic complexity measurements included or alluded to unit length, sentence complexity, subordination indices, coordination, specific structures, working memory load, referring expressions, syntactic simplicity, similarity, and variation. For each CEFR competence level, MANOVA was used to compare the different age groups.

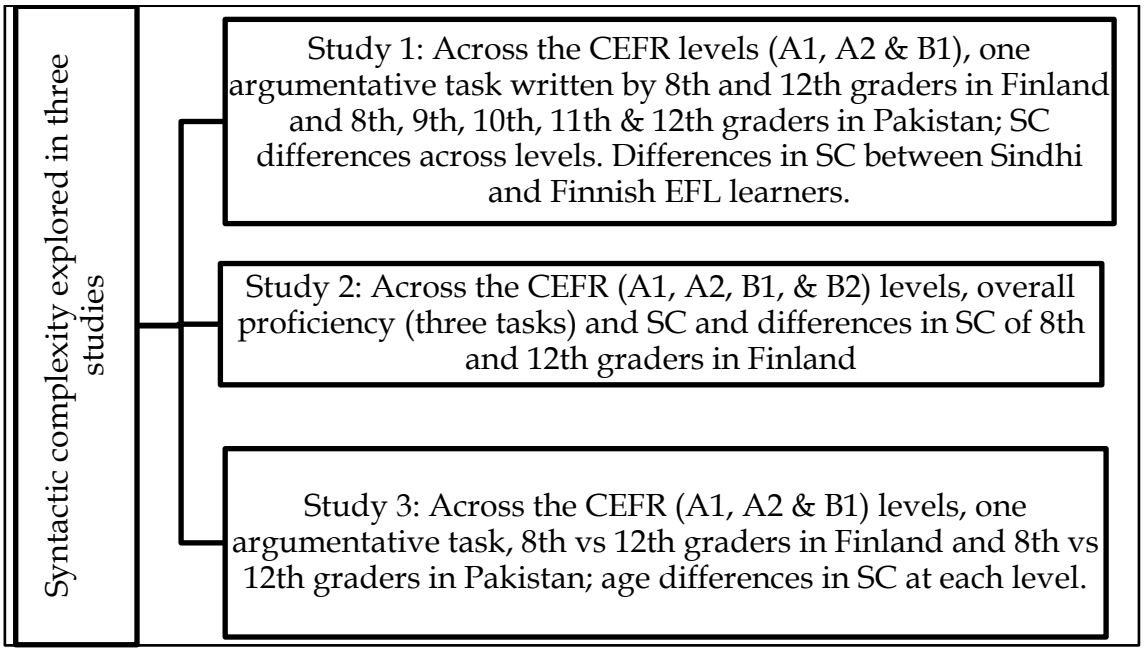
The results demonstrated that while most syntactic complexity elements remained constant regardless of the overall CEFR level (in writing), several features differed between the two age groups. The findings suggest that language learners' age (and level of education, which usually go hand in hand with (young) language learners' age) can influence some of the syntactic complexity characteristics at certain competence levels. At the A1 level, practically all SC indices among Pakistani learners showed an age effect, whereas just a few indices at the A2 and B1 levels did. Overall, the age effect was very substantial in Pakistan. In contrast, few significant differences in syntactic complexity features existed between eighth- and twelfth grade EFL learners in Finland at the A2 and B1 levels. In summary, the Pakistani EFL learners were more varied than their Finnish counterparts at a single CEFR level because of age.

Finally, a brief description is in order of the division of work between the authors in the two co-authored articles (based on Study 1 and Study 2) that form part of the doctoral dissertation of Ghulam Abbas Khushik. In Studies 1 and 2, the first author (i.e., Khushik) carried out data gathering and rater training in Pakistan and organised the transcription of the handwritten scripts. He was also responsible for cleaning the data for automated analyses of SC (including the Finnish data that had not been cleaned previously because automated analyses of the Finnish learners' texts had not been done in the Dialuki project). He conducted all the statistical and linguistic analyses of Pakistani and Finnish data; Facets analyses were jointly conducted with Professor Huhta. Regarding the writing of the articles, the first author had primary responsibility for all the sections in them. In contrast, the second author (i.e., Huhta) focused on finalising matters related to the style and language of the articles and provided input to specific content points, such as information about the Finnish context and data gathering about the performances from the Finnish participants.

9 DISCUSSION

This dissertation examined differences in syntactic complexity features across the CEFR levels among EFL learners from two countries and different age groups. To address this aim, I collected written data from Pakistan and used existing data from Finland. The data consisted of narrative, descriptive, and explanatory essays. The essays were rated at the six CEFR levels. The handwritten scripts were transcribed to analyze them linguistically using automated computer tools. The linguistic analysis provided the basis for the statistical analyses to determine if the CEFR levels differed regarding various syntactic complexity features and if the differences were similar in the two contexts (Pakistan and Finland) and different age groups in each country. Overall, the findings indicate that the CEFR levels could be distinguished linguistically via syntactic complexity features and that there were some notable differences in the two contexts and age groups.

Figure 9.1 Key points of the three studies in the dissertation



Further discussion is continued in several sections. First, I will discuss the findings relating to the differences in SC between CEFR levels based on Study 1 and Study 2 (i.e., Article 1 and Article 2, respectively). This discussion is divided into two parts. Section 9.1 and its subsections account for the extent to which different groups (types) of SC indices are separated across the CEFR levels covered in the two studies. Special attention is paid to the effect sizes in reporting these findings since they indicate how significant the identified differences were (statistical significance does not always indicate this). Each sub-section of 9.1 discusses the similarities and differences between the results of studies 1 and 2 and previous CEFR-related research on EFL learners that have included the SC indices. The second significant section, 9.2, summarizes the findings of Studies 1 and 2 concerning pairs of CEFR levels. Thus, it gives an overview and a list (in a table format) of the SC indices that significantly differentiated (1) between A1 and A2, (2) between A2 and B1, and (3) between B1 and B2, as well as (4) between the non-adjacent levels A1 and B1.

The third significant section in the discussion, Section 9.3, focuses on the age of the learners and whether syntactic complexity seemed to be related to the learners' age. Study 2 included two EFL learner groups from Finland who differed in age (14-year-old 8th graders and 17-year-old upper secondary school students). Although Study 2 did not directly compare the two age groups (e.g., by using statistical analyses such as t-tests), the differences in the results between the two groups suggest some age-related differences in their use of syntactic structures in their EFL writing. Details of these differences were reported in Article 2 and in Section 9.1 of the Discussion. Learners' age was, however, the focus of Study 3 (i.e., Article 3), which was based on two age groups (8th and 12th grade) in two countries (Pakistan and Finland) and explicitly investigated if the learners whose CEFR writing level was the same but whose age differed used equally complex syntax.

It should be noted that the learners' CEFR writing level was defined somewhat differently in the three studies. However, the CEFR level was always based on the same writing task or tasks in both countries (when learners from two countries were investigated, as in Studies 1 and 3). In Study 1, the learners' level was based on their performance on one argumentative essay. In Study 2, on Finnish EFL learners, the learners' level was based on three writing tasks, one of which was the same in both age groups. Study 3 used the same writing task as Study 1. However, the focus was different (comparing SC in the writing of learners whose age differed but whose CEFR level was the same, rather than comparing SC at different CEFR levels and L1 groups, as was the case in Study 1).

To recap some of the earlier discussion of the key definitions in this dissertation, "complexity" refers to the availability and accessibility of primary and sophisticated structures in language production. Due to the absence of such complexity, the range of available and accessible structures in language production can be considered somewhat limited (Wolfe-Quintero, 1998).

Syntactic complexity refers to the variety of forms and the sophistication of such forms in linguistic production (Ortega, 2003). These definitions imply that the presence (or lack) of such patterns and structures in learners' works shows their proficiency. The exploration of several indices and the study of such forms qualitatively and quantitatively can be used to confirm the learners' competency levels and language progress. The indices' measurements provide information about linguistic aspects such as the sophistication and simplicity of the language.

I agree with Bulté and Housen's (2014) view that later-developed syntactic complexity features or later-acquired linguistic features indicate the developmental process of language from beginner to intermediate and from intermediate to advanced level. It could be said that later acquired or developed features are correlated with higher proficiency levels. The second idea is that more proficient learners use more complex language, meaning they are more advanced, complicated, mature, and developed. Indeed, I found from the dissertation that most of the syntactic complexity features increase to make the language more complex at higher proficiency levels. However, the values of certain features (e.g., indices related to syntactic similarity and simplicity) decrease to make the language more complex. Therefore, it is essential to consider all syntactic complexity features to see how they increase or decrease in the current dissertation to understand the more complex language presented by Bulté and Housen (2014).

The general findings of this dissertation indicate that many syntactic complexity features extracted from written tasks significantly differentiated the CEFR levels in the Pakistani and Finnish EFL learners' writing. Similarly, syntactic complexity features presented significant differences due to the different L1s in the two countries (Pakistani Sindhi and Finnish, in Study 1) and due to the different ages of the learners (Finland, in Study 2, and particularly in Pakistan, in Study 3).

9.1 The current dissertation's findings on differences in SC across CEFR levels in comparison with previous studies

This dissertation examined the CEFR's linguistic foundations, which have been considered significant given the Framework's prominence (e.g., Hulstijn 2007; Bartning et al., 2010; Wisniewski, 2017). While numerous studies on Second Language Acquisition (SLA) have investigated the relationship between linguistic qualities and language competency, few have examined language proficiency in terms of the CEFR levels and, consequently, the linguistic characteristics associated with those levels or those that distinguish between the levels. However, it is essential to note that the current study investigated syntactic complexity differences (and similarities) between the CEFR levels by using and referring to the numerical values of the various SC indices. Which

syntactic structures might characterise EFL writing at CEFR levels among the Sindhi and Finnish EFL learners were outside the scope of the current study.

This study aimed to evaluate whether syntactic complexity indices might be used to distinguish between CEFR levels in EFL learners' writings. Furthermore, the SC features of groups of English-language learners from Finland and Pakistan were compared with shedding light on the generalizability of the findings across EFL learners with different L1 backgrounds and ages. In this section, I will first summarise the findings of Study 1 and Study 2 as they both focused on investigating differences in SC across different CEFR levels (even though Study 2 also compared two age groups covered in a more detailed fashion in Study 3). After summarising the findings of this dissertation, the findings will be compared with earlier studies on SC in EFL students' writings by focusing on the SC characteristics that differentiate the CEFR levels. Prior research on the linguistic differences between the CEFR levels was limited, and those conducted used somewhat different sets of SC indicators. As a result, a comprehensive picture of the SC qualities that distinguish the CEFR levels in EFL students' writing is difficult to obtain. However, by combining this dissertation's results and previous studies' findings, we can better understand how SC differentiates between CEFR levels.

Table 3 in Appendix 1 below, as well as Table 1 in Study 1 and Table 14 in Study 2, provide summaries of the findings regarding SC indices that have been found to significantly differentiate between various CEFR levels, both in the studies comprising this dissertation and studies carried out by other researchers.

9.1.1 Length of the production units

Prior to the introduction of the CEFR, Wolfe-Quintero et al. (1998, p. 23), for example, found that the length of sentences increases with proficiency. This appears to be true for several levels of proficiency adjacent to each other. Numerous studies related to the CEFR have investigated the relationship between different production units and proficiency levels. This is similar to the way that as a child grows up, they continue to add more words to their vocabulary. As they gain more knowledge and experience, they are better able to express themselves with longer and more complex sentences.

The current study found that the Finnish and Sindhi groups have differences in *mean sentence length, T-unit length,* and *clause length* across the CEFR levels A1–B1/B2. The Sindhi-speaking EFL learners consistently wrote longer clauses, T-units, and sentences than their Finnish peers with the same CEFR writing level (see Study 1, p. 517).

Appendix 3 includes some examples of texts written by Sindhi and Finnish EFL learners that illustrate, for example, differences in these SC indices. The sentences' length and T-units were the only SC variables that distinguished all the CEFR levels A1-B1 in both groups. The length of production units also increased with proficiency in Study 2, which focused on Finnish EFL learners. Again, sentence length and, to some extent also, T-unit length were quite clear separators of the CEFR levels. The standard deviation of sentence length also

separated the CEFR levels: learners with lower proficiency mostly wrote sentences whose length did not vary much. Higher-proficiency learners showed more variation in their sentences, writing shorter and longer sentences. Interestingly, the length of clauses did not separate the levels in either L1 group in Study 1 (apart from Finnish B1s from A1/A2s). In Study 2 on Finnish EFL learners, clause length was, however, a clear separator of B1 vs. B2 learners (in that study, it was possible to investigate learners up to the B2 level, particularly in the older gymnasium student group).

To summarise, in Study 1, comparing Sindhi and Finnish EFL learners at CEFR levels A1 to B1, the following findings were obtained for the length of production units (see Tables 5 and 9 in Supplementary Data and Figure 1 in Study 1):

- *The mean length of sentence*: a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 and the adjacent levels A1 vs A2 and A2 vs B1 in both groups; the Sindhi wrote longer sentences than the Finns with the same overall CEFR writing level
- *The standard deviation of mean length of sentence*: a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 and A1 vs A2 in both L1 groups but for A2 vs B1 only for the Finnish group; the standard deviations were higher for the Sindhi than for the Finns
- *The mean length of T-unit*: a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 and A2 vs B1 in both L1 groups but for A1 vs A2 only for the Sindhi group; the Sindhi wrote longer T-units than the Finns.
- *The mean length of the clause*: significant overall increase only for the Finnish group; significant pairwise differences for A1 vs B1 and A2 vs B1 in the Finnish group; the Sindhi wrote longer clauses than the Finns

The effect sizes of the overall differences between CEFR levels (from A1 to B1) were somewhat larger for the Finnish group. For them, the largest effect sizes were found for the mean sentence length ($\eta 2$= 0.248), the standard deviation of the sentence length ($\eta 2$= 0.186) and the mean T-unit length ($\eta 2$= 0.104). These can all be considered significant or at least medium effect sizes if we consider 0.01 small, 0.06 medium and 0.14 (or higher) large effect sizes (see, e.g., Richardson's 2011 evaluation supporting Cohen's original (1969) estimates of small-medium-large effect sizes for partial eta squared values reported in analyses of variance). None of the effect sizes for the Sindhi group exceeded 0.6, the largest for the mean sentence length ($\eta 2$= 0.056) and the mean T-unit length ($\eta 2$= 0.041); these were, thus, small effect sizes.

In Study 2, focusing on two Finnish groups of EFL learners (8th-graders and 12th-grade gymnasium students at CEFR levels A1 to B2), the following findings were obtained for the length of production units (see Tables 10, 11, 12 and 13, and Figure 3 in Study 2):

- *The mean length of sentence*: a significant overall increase in both age groups; significant pairwise differences only for A1 vs A2 in the 8th grade; the gymnasium students wrote longer sentences than the 8th graders with the same overall CEFR writing level
- *The Standard deviation of mean length of sentence*: a significant overall increase in both age groups; significant pairwise differences for A1 vs A2 and A2 vs B1 in the 8th grade (A2 vs B1 approached significance in the gymnasium); the standard deviations were the same in both age groups for learners with the same CEFR level
- *The Mean length of T-unit*: a significant overall increase in both age groups; significant pairwise differences only for A1 vs A2 in the 8th grade; the gymnasium students wrote longer T-units than the 8th graders
- *The Mean length of the clause*: a significant overall increase in the gymnasium group (almost significant for the 8th graders); significant pairwise differences only for B1 vs B2 in the gymnasium group; the gymnasium students wrote longer clauses than the 8th graders

The effect sizes of the differences between CEFR levels were larger for the 8th graders. Three large effect sizes for this group were found: the standard deviation of sentence length ($\eta 2$= 0.24), mean sentence length ($\eta 2$= 0.20), and mean T-unit length ($\eta 2$= 0.14). For the gymnasium group, the only effect size exceeding the threshold for medium effect size was for the mean clause length ($\eta 2$= 0.11).

The results of the current dissertation are consistent with the findings of much of previous CEFR-related research concerning production unit lengths (see Table 3 in Appendix 1). Although Lu and Ai (2015, p. 22) did not use the CEFR as the reference point in their study of SC, it is of interest report their findings regarding the mean length of sentences, clauses, and T-units because they compared eight different L1 groups. According to their findings, the Japanese L1 learners used 14.6 words per sentence. There was a gradual increase in the number of words per sentence for the Chinese, Russians, Bulgarians, French, and Tswanas, and it reached 22.3 words per sentence for the Germans. Their results are in line with the finding of this dissertation (e.g., Study 1) as they, too, showed that learners of different L1 backgrounds tend to construct their sentences in English differently which results in differences in sentence length. There was a similar gradual increase in the length of T-units and clauses for Japanese, Chinese, Russians, Bulgarians, French, Tswanas, and Germans.

The English Profile-based studies reported by Hawkins and Filipovic (2012) discovered that the *length of a sentence* distinguished between the CEFR levels from A2 to C2. Also, Alexopoulou et al. (2017) found sentence length to separate EFL learners in the A1-B2 range, and Lahuerta Martínez (2018) discovered it to distinguish between A2 and B1 learners, as did Polat et al. (2020) and Barrot and Agdeppa (2021) for A2 vs B1 and B1 vs B2. Regarding *T-unit length*, it has been demonstrated by Gyllstad et al. (2014) and Verspoor et al. (2012) that it can be utilised to distinguish between A1 and A2, as well as between A2 and B1. Both Polat et al. (2020) and Barrot and Agdeppa (2021) found T-unit length to separate not only A2s from B1s but also between B1 and B2 (see Appendix 1, Table 2). *Clause length* seems to have been examined somewhat less often than sentence and T-unit lengths in previous CEFR-related research. However, unlike most of the findings in this dissertation, those studies have found some significant results. Gyllstad et al. (2014) found clause length to separate A2 from B1, and Alexopoulou et al. (2017) found all the A2–B2 range levels. Polat et al. (2020) and Barrot and Agdeppa (2021) found clause length to distinguish between A2 and B1 and B1 and B2. Finally, Paquot's (2019) results showed that clause length increased and separated the three top CEFR levels of B1, C1 and C2 from one another.

9.1.2 Sentence complexity, subordination, and coordination

The links between the sentence's primary sub-sentential pieces (clauses and T-units) were also investigated with reference to sentence complexity, subordination, and coordination. In Study 1 (see Figures 2 and 3 in Study 1), the complexity of sentences (*the number of clauses per sentence* and the *number of T-units per sentence*) increased somewhat from one CEFR level to another. Still, only A1 vs. A2 was significant in both L1 groups (however, for the Sindhi group, T-units per sentence were not significant). Both groups wrote similar clauses per sentence, but the Sindhi EFL learners included significantly more T-units in their sentences. The two L1 groups did not differ from each other. The findings from Study 2 on the Finnish EFL learners were quite like those obtained for the Finnish learners in Study 1.

The four indices of subordination (*dependent clauses per clause and T-unit, complex T-units per T-unit,* and *clauses per T-unit*) increased from level to level in both L1 groups in Study 1 (Figure 2 in Study 1). However, the picture could have been clearer for Study 2, particularly for the older gymnasium group of Finnish EFL learners (Figure 4 in Study 2). Dependent clauses per clause and dependent clauses per T-unit increased significantly (or almost significantly) across A1 and B1 in both L1 groups in Study 1. However, in Study 2 on Finnish EFL learners, those two indices of subordination only distinguished A1 vs A2 learners and only in the 8th graders' group. Regarding the other two indices of subordination (complex T-units per T-unit and clauses per T-unit), in Study 1, they increased significantly from A1 to A2 in the Sindhi group. However, only the complex T-units per T-unit did that for the Finns and only for A2 vs B1. Study 2, focusing on the Finnish learners, yielded similar results in that there were no significant

differences. In contrast to Study 1, clauses per T-unit separated A1 from A2 in the 8th graders' group.

The two indices of syntactic coordination (*coordinate phrases per clause* and *coordinate phrases per T-unit*) did not distinguish between any of the CEFR levels investigated in either Study 1 or Study 2. Neither could any clear and consistent trends (upwards or downwards) be seen in the development of these indices across CEFR levels (apart from a possible slight increase among the Finnish learners in Study 1; see Figure 3 in Study 1 and Figure 5 in Study 2). However, the two L1 groups differed because the Sindhi speakers wrote more coordinated phrases per clause and T-unit than their Finnish peers (see Figure 3 in Study 1).

To summarize, in Study 1, comparing the Sindhi and Finnish EFL learners at CEFR levels A1-B1, the following findings were obtained for the SC indices relating to sentence complexity, subordination, and coordination (see Tables 6 and 10 in Supplementary data and Figures 2 and 3 in Study 1):

- *Number of clauses per sentence*: a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 and A1 vs A2 in both L1 groups; both L1 groups used roughly the same number of clauses per sentence.
- *Number of T-units per sentence:* a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 in both L1 groups, and for the Sindhi group also for A1 vs A2 (approached significance for the Finns, too); the Sindhi used more T-units per sentence than their Finnish peers.
- *Dependent clauses per clause:* a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 in both L1 groups, and for A2 vs B1 (although marginally significant for the Sindhi); for A1 vs A2, marginally significant for the Sindhi and approached significance for the Finns; the Finns used more dependent clauses per clause than their Sindhi peers.
- *Dependent clauses per T-unit:* a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 in both L1 groups, and for A1 vs A2 in the Sindhi group; A1 vs A2 for the Finns and A2 vs B1 in both L1 groups approached significance; the Finns used more dependent clauses per T-unit than their Sindhi peers.
- *Complex T-units per T-unit:* a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 in both L1 groups; significant for A2 vs B1 in the Finnish group and marginally significant in the Sindhi group; the Finns used more dependent clauses per T-unit than their Sindhi peers.
- *Clauses per T-unit:* a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 in both L1 groups;

significant for A1 vs A2 in the Sindhi group (approached significance in the Finnish group); the Finns used more clauses per T-unit than their Sindhi peers.

- *Coordinate phrases per clause:* no clear trends across CEFR levels in either of the L1 groups and no significant differences; the Sindhi used more coordinate phrases per clause than their Finnish peers.
- *Coordinate phrases per T-unit:* slight but insignificant increase in the Finnish group and no clear trend in the Sindhi group; the Sindhi used more coordinate phrases per T-unit than their Finnish peers.

The effect sizes of the overall differences between CEFR levels (A1 to B1) were larger for the Finnish group. However, only medium effect sizes were found for dependent clauses per clause ($\eta 2$= 0.097), complex T-units per T-unit ($\eta 2$= 0.094), clauses per sentence ($\eta 2$= 0.084), and dependent clauses per T-unit ($\eta 2$= 0.064). For the Sindhi group, only some tiny effect sizes (i.e., between 0.01 and 0.06) were found, the largest being for clauses per T-unit ($\eta 2$= 0.029) and dependent clauses per T-unit ($\eta 2$= 0.025) and dependent clauses per clause ($\eta 2$= 0.021).

In Study 2, focusing on two Finnish groups of EFL learners (8th graders and gymnasium students at CEFR levels A1-B2), the following findings were obtained for the SC indices relating to sentence complexity, subordination, and coordination (see Tables 10 and 11 and Figures 3, 4 and 5 in Study 2):

- *Number of clauses per sentence*: a significant overall increase in only the 8th-grade group; the significant pairwise difference for A1 vs A2 in the 8th grade; both age groups used roughly the same number of clauses per sentence
- *Number of T-units per sentence:* overall increase approaching significance in the 8th-grade group; no significant pairwise differences; both age groups used roughly the same number of T-units per sentence.
- *Dependent clauses per clause:* overall non-significant increase in both age groups (approached significance in the 8th grade); the almost significant pairwise difference for A1 vs A2 in the 8th grade; the gymnasium group used more dependent clauses per clause than the 8th graders.
- *Dependent clauses per T-unit:* overall non-significant increase in both age groups (approached significance in the 8th grade); the almost significant pairwise difference for A1 vs A2 in the 8th grade; both age groups used roughly the same number of dependent clauses per T-unit.
- *Complex T-units per T-unit:* no clear trends in either age group across CEFR levels; the gymnasium group used more complex T-units per T-unit than the 8th graders.

- *Clauses per T-unit:* a significant overall increase in only the 8th-grade group; the significant pairwise difference for A1 vs A2 in the 8th grade; both age groups used roughly the same number of clauses per T-unit
- *Coordinate phrases per clause:* no clear trends in either age group across CEFR levels; the gymnasium group used more coordinate phrases per clause than the 8th graders.
- *Coordinate phrases per T-unit:* no clear trends in either age group across CEFR levels, but almost significant overall differences and a pairwise increase from A1 to A2 in the 8th grade; the gymnasium group used more coordinate phrases per T-unit than the 8th graders.

The effect sizes of the overall differences between CEFR levels (A1 to B2) in the two Finnish age groups were medium at best and only in the 8th graders' group. No statistically significant overall or pairwise differences were found for the gymnasium group, so almost all effect sizes were below 0.01. The medium effect sizes observed in the 8th grade included clauses per sentence ($\eta 2$ = 0.09), clauses per T-unit ($\eta 2$= 0.08), dependent clauses per clause ($\eta 2$ = 0.06), and T-units per sentence ($\eta 2$= 0.06).

Previous CEFR-related studies on syntactic complexity have addressed the complexity of sentences, T-units, and clauses. Sentence-level complexity has received less attention than the others, but some significant differences between CEFR levels have emerged. In contrast to the results of the current dissertation, previous research has found clauses per sentence to separate EFL learners in the A2 to B2 range. Polat et al. (2020) found it to separate A2 from B1 and B1 from B2, and Barrot and Agdeppa (2021) found it to distinguish between A2 and high B1 (or B1.2) learners but also within B1, that is, between low (B1.1) and high (B1.2) B1 learners. Barrot and Agdeppa also examined T-units per sentence but found no significant results between CEFR levels. Lahuerta Martinez (2018) also found the ratio of complex sentences to separate A2 and B1 level learners (she defined a *complex sentence* as a sentence that includes at least one dependent clause, so sentence level complexity in her study may not be the same as in these other studies). Since none of these previous studies included A1 level learners, whom I found to differ from A2 learners in terms of clauses per sentence and, in the Finnish L1 group, also in terms of T-units per sentence, it is somewhat difficult to form a clear picture of how well sentence level complexity measured as clause or T-unit ratios distinguishes between CEFR levels. However, the available evidence suggests that clauses per sentence, at least a somewhat consistent separator in the A1 to B2 range for EFL writing.

The findings of current dissertation align with Lu and Ai (2015) study for most of the measures. For example, the current dissertation reports that there were differences between Sindhis and Finns for *dependent clauses per clause and T-unit, complex T-units per T-unit,* and *clauses per T-unit, coordinate phrases per clause* and *coordinate phrases per T-unit*. In their study of eight groups of different

L1 English learners, Lu and Ai (2015) found that dependent clauses per clause and T-unit differed, as did complex T-units per T-unit, clauses per T-unit, and coordinate phrases per clause and coordinate phrases per T-unit. For the Japanese, Chinese, Russians, Bulgarians, French, Tswanas, and Germans, there was a similar gradual increase in dependent clauses per clause and T-unit, complex T-units per T-unit, and clauses per T-unit. In all four variables, the Japanese had the lowest mean score, while the Germans had the highest mean scores. This suggests that, while all of these L1 languages affected the learners' syntax in the English language, there were still differences in the complexity of their syntax. For example, the L1 Japanese EFL learners had fewer dependent clauses per clause and complex T-units per T-unit than the other languages, while the Germans had more.

Previous research has extensively studied subordination at the clause and T-unit levels. For example, Wolfe-Quintero et al. (1998) argued that dependent clauses per clause are a good proxy for linguistic proficiency. Regarding the four indices of subordination included in the current study (*dependent clauses per clause, dependent clauses per T-unit, complex T-unit per T-unit,* and *clauses per T-unit)*, the following results have been obtained in previous CEFR-related research. In the early CEFR-related study by Kim (2004), clauses and dependent clauses per T-unit and dependent clauses per clause were found to distinguish between B1 and B2 and somewhat less clearly between A2 and B1. Gyllstad et al. (2012) found clauses per T-unit to separate A2 from B1. Alexopoulou et al. (2017) discovered that subordinate clauses per T-unit separated all adjacent levels in the A1 to B2 range. Lahuerta Martinez (2018) found a dependent clause ratio to distinguish between A2 and B1 learners. Polat et al. (2020) studied all four abovementioned subordination indices and found them to separate both A2 and B1 and B1 and B2. Also, Barrot and Agdeppa (2021) covered all these indices in their study and found similar results as Polat et al. (2020) in the A2–B2 range (also between low- vs high-level B1 learners). Finally, Paquot (2019) investigated three indices (clauses per T-unit, dependent clauses per T-unit, dependent clauses per clause) with advanced EFL learners and found them to separate B2 from C1 and C1 from C2 learners.

Subordination at the clause and T-unit level has, thus, been found to be a consistent distinguishing feature across the whole CEFR scale range. Although the findings of the current dissertation included some non-significant results, these findings were in line with previous research in that subordination indices were good separators of CEFR levels among the learners involved in the current research study.

In previous CEFR research, coordination at the clause and T-unit level (operationalized as *coordinate phrases per clause* and *coordinate phrases per T-unit*) has been studied less often than subordination. However, some recent studies have found some significant results for coordination indices. Lahuerta Martinez (2018) found a coordinate clause ratio to distinguish A2 from B1 learners. Polat et al. (2020) discovered that coordinating phrases per clause and T-unit differentiated A2 vs. B1 and B1 vs. B2 learners. Similar results were also obtained

by Barrot and Agdeppa (2021), who found the two indices to separate A2 learners from high B1 (and B2) learners and low B1 learners from B2 learners, as well as low vs. high B1 learners (i.e., B1.1 vs. B1.2). Thus, these findings differ from the non-significant results obtained in the current dissertation, where the coordination indices did not separate learners at different CEFR levels.

9.1.3 Degree of phrasal sophistication

The number of *complex nominals in a sentence*, the number of *complex nominals in a T-unit*, and the number of *verb phrases in a T-unit* are elements of phrasal sophistication. They are, thus, essential features of observing the syntax across the CEFR levels. The verb phrases per T-unit are essential, but perspectives on their usefulness differ. Wolfe-Quintero et al. (1998: 85–123) recommend verb phrases per T-unit since they contribute to the overall measurement of SC. However, in his non-CEFR study, Lu (2011) did not find this index to separate the different proficiency groups.

In the current dissertation, *verb phrases per T-unit* were a reasonably good separator of CEFR levels. In Study 1, A1 and B1 learners could be very clearly distinguished from one another in both the Sindhi and Finnish L1 groups. Also, the adjacent CEFR levels A1 vs. A2 and A2 vs. B1 significantly differed in both L1 groups. In Study 2, however, focusing on the Finnish learners, only A1 and A2 learners in the 8th graders' group were separable, but no significant differences were found for the gymnasium students. Comparing the two L1 groups in Study 1 (see Figure 4 in Study 1) also shows that the Sindhi-speaking EFL learners used fewer verb phrases per T-unit than their Finnish peers.

Significant differences were found between A1 and B1 learners in Study 1 in both L1 groups for complex nominals per T-unit. However, the adjacent CEFR levels could not be distinguished, although the A2 vs. B1 difference in the Sindhi group was marginally significant. No significant findings were obtained in Study 2 for Finnish EFL learners. For *complex nominals per clause*, no significant differences were found for either of the L1 groups in Study 1. However, in Study 2, the B1 and B2 gymnasium students could be separated from one another concerning this index of phrasal sophistication. Again, an L1 effect could be seen in the results (see Figure 4 in Study 1): the Sindhi-speaking EFL learners used more complex nominals per clause and T-unit than their Finnish peers (note that the situation was the reverse for the number of verb phrases per T-unit, as mentioned above).

To summarize, in Study 1, the following findings were obtained for the SC indices relating to phrasal sophistication (see Table 6 in Supplementary data and Figure 4 in Study 1):

- *Verb phrases per T-unit:* a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 and the adjacent levels A1 vs A2 and A2 vs B1 in both groups; the Finns

wrote more verb phrases per T-unit than the Sindhi with the same overall CEFR writing level

- *Complex nominals per clause:* no clear trends across CEFR levels in either of the L1 groups (although there was a slight increase from A2 to B1 in the Finnish group and an almost significant overall increase in that group); the Sindhi used more complex nominals per clause than the Finns
- *Complex nominals per T-unit:* a significant overall increase in both L1 groups; significant pairwise differences for A1 vs B1 in both L1 groups and differences approaching significance for A2 vs B1 in both groups; the Sindhi used more complex nominals per T-unit than the Finns

The effect sizes of the overall differences between CEFR levels (A1 to B1) were larger for the Finnish EFL learners, with one large effect size ($\eta 2$= 0.14) found for verb phrases per T-unit and one medium effect size for complex nominals per T-unit ($\eta 2$= 0.061). The largest effect sizes (though small, i.e., under 0.06) for the Sindhi group were found for the same two SC indices: verb phrases per T-unit ($\eta 2$= 0.038) and complex nominals per T-unit ($\eta 2$= 0.019).

In Study 2, focusing on two Finnish groups of EFL learners (8th graders and gymnasium students at CEFR levels A1-B2), the following findings were obtained for the SC indices relating to phrasal sophistication (see Tables 10 and 11 and Figure 6 in Study 2):

- *Verb phrases per T-unit:* significant overall increase for the 8th graders; the significant pairwise difference for A1 vs A2 for the 8th graders; the gymnasium group used more verb phrases per T-unit than the 8th graders
- *Complex nominals per clause:* significant overall increase for the gymnasium students; the significant pairwise difference for A2 vs B1 for the gymnasium group; the gymnasium group used more complex nominals per clause than the 8th graders
- *Complex nominals per T-unit:* overall increase approaching significance for the gymnasium students; the almost significant pairwise difference for A2 vs B1 for the gymnasium group; the gymnasium group used more complex nominals per clause than the 8th graders

The effect sizes of the overall differences between CEFR levels (A1 to B2) among the Finnish EFL learners were roughly similar in the two age groups, but the pattern was somewhat different. A medium effect size was found in the 8th-grade group for verb phrases per T-unit ($\eta 2$= 0.10) but in the gymnasium group

for complex nominals per clause ($\eta 2$= 0.07). Similar, small ($\eta 2$= 0.03) effect sizes were found in both age groups for complex nominals per T-unit.

According to the current dissertation, Sindhi and Finnish L1 EFL learners differed in the number of complex nominals per clause and T-unit, verb phrases per T-unit, and clauses per sentence. The findings of this study are consistent with those of Lu and Ai (2015). According to Lu and Ai (2015), they observed differences in the number of complex nominals per clause and T-unit, verb phrases per T-unit, and clauses per sentence across eight groups of different L1 English learners. In all variables, the Japanese, Chinese, Russians, Bulgarians, French, Tswanas, and Germans exhibited different behavior. Each variable had a different mean score between the L1 with the lowest mean score and the L1 with the highest mean score. It was therefore evident that, even though all of these L1 languages influenced a learner's syntax in English, there were still differences in the complexity of their syntax. This is likely due to the fact that each language has its own unique grammatical structure and rules, which can influence the way a learner approaches and acquires English. This could explain the differences between the mean scores between the different L1s.

Indicators of phrasal sophistication have received some attention in the more recent CEFR-related research. Regarding the number of *verb phrases per T-unit*, both Paquot (2019) and Polat et al. (2020) found significant differences between learners in the CEFR scale range A2–C2 (Polat et al. for A2 vs. B1 and B1 vs. B2, and Paquot for B2 vs. C1 and C1 vs C2). Barrot and Agdeppa (2021), too, found the index to distinguish between A2 and high B1 (and B2), within B1 (low vs. high B1), and between low/high B1 and B2. Regarding the *number of complex nominals per clause and complex nominals per T-unit*, the three studies mentioned above yielded almost precisely the same results as those they found for the verb phrases. Polat et al. (2020) found both indices to separate A2 vs B1 and B1 vs B2, and Paquot (2019) found B2 vs C1 and C1 vs C2. Barrot and Agdeppa's (2021) findings for the two phrasal indices based on complex nominals were even more evident than those they obtained for the verb phrases (the effect sizes were larger; see Table 6 in their article): all the adjacent and non-adjacent proficiency groups involved in their study could be significantly distinguished from one another, including A2 vs low B1. In addition, two other CEFR-related studies, while not focusing on complex nominals, can be mentioned here. Kim (2004) found the number of nominals per clause to be one of the significant separators of A2 and B1 levels, and Lahuerta Martinez (2018) found noun phrases per clause to separate those same CEFR levels.

The results of the current dissertation studies and those obtained by previous investigations based on the CEFR levels are relatively compatible, particularly regarding the number of verb phrases per T-unit. This SC index is a consistent separator across the entire CEFR scale. The results relating to the complex nominals in clauses and T-units were somewhat mixed. Both this dissertation and previous studies show the number of complex nominals per T-unit to be a rather suitable separator of CEFR levels. However, unlike in other studies, the number of complex nominals per clause did not distinguish

proficiency levels in the current research. A possible reason for this difference might be that the mean length of clauses did not differ significantly between adjacent CEFR levels in either Study 1 or 2; the T-unit and sentence length did, at least in most comparisons of levels.

9.1.4 Verb and noun phrases

Two indices of verb and noun phrases available in Coh-Metrix can be used to assess the complexity of syntactic structures at the phrasal level. The first is called *the left embeddedness* of a sentence, and it is determined by the number of words that precede the verb in the primary phrase of the sentence. When more words occur before the verb, sentences densify and are complicated, which taxes working memory and makes such constructions more challenging to process (Graesser et al., 2004), notably the weaker the learner's proficiency is. This second index considers the number of *modifiers per noun phrase*, which relates to how difficult it is to refer to assertions in a text (Graesser et al., 2004).

The findings concerning left embeddedness in Study 1 of this dissertation show that the values of this index rose steadily across CEFR levels and were consistently higher for the Sindhi-speaking EFL learners compared with their Finnish peers (see Figure 5 in Study 1). However, significant differences were found only for the Sindhi group, where A1 and B1 learners could be separated, and A1 vs. A2 was also marginally significant. Left embeddedness did not separate Finnish learners in either Study 1 or 2. The findings concerning the number of modifiers per noun phrase were different in at least two ways: first, significant differences between CEFR levels were found only for the Finnish EFL learners, and second, those findings suggest non-linear development for this SC index at the first three or four of the CEFR levels. In Study 1, significant differences were found for the Finnish learners' A1 vs. A2 and A2 vs. B1 levels (see Figure 5 in Study 1). As with left embeddedness, the two L1 groups differed because the Sindhi group used more modifiers per noun phrase than the Finns. In Study 2, a significant difference was also found for B1 vs. B2 in the Finnish gymnasium group (see Figure 8 in Study 2). The difference between A2 and B1 was also considerable, even if it did not quite reach significance. The more exciting finding concerning the number of modifiers per noun phrase was that the significant differences found in Study 1 for the Finnish L1 group indicated non-linear development: the values first decreased from A1 to A2, then increased from A2 to B1. The results from Study 2 for the gymnasium group showed that the values continued rising significantly from B1 to B2. For Pakistani learners, the trend was similar in that the values first decreased and then increased. However, the changes from one level to the next were insignificant (see also the discussion about noun phrase density below).

To summarize, in Study 1, comparing Sindhi and Finnish EFL learners, the following findings were obtained for the SC indices relating to verb and noun phrases (see Tables 7 and 11 in Supplementary data and Figure 5 in Study 1):

- *Left embeddedness:* an overall increase in both L1 groups but significant only in the Sindhi group; the significant pairwise difference for A1 vs B1 and (marginally significant) for A1 vs A2 in the Sindhi group; the Sindhi used longer left embedding than their Finnish peers with the same CEFR level
- *Modifiers per noun phrase*: non-linear trend in the Finnish group, decreasing significantly from A1 to A2 and then increasing significantly from A2 to B1; no trend in the Sindhi group; the Sindhi wrote more modifiers per noun phrase than the Finns

The effect sizes of the overall differences between CEFR levels (A1 to B1) were small or very small for these SC indices. The largest was for left embeddedness in the Sindhi group ($\eta 2$ = 0.019).

In Study 2, focusing on two Finnish groups of EFL learners (8th graders and gymnasium students at CEFR levels A1-B2), the following findings were obtained for the SC indices relating to verb and noun phrases (see Tables 12 and 13 and Figure 8 in Study 2):

- *Left embeddedness:* an overall increase in both age groups (significant in the gymnasium group and almost significant in the 8th grade); the significant pairwise difference for B1 vs B2 in the gymnasium group and almost significant for A1 vs A2 in the 8th grade; the gymnasium students used more left embeddedness than the 8th graders
- *Modifiers per noun phrase:* a significant overall increase in the gymnasium group (possible slight but non-significant non-linear trend like Study 1 in the 8th grade); the significant pairwise difference for B1 vs B2 in the gymnasium group; the gymnasium students used more modifiers per noun phrase than the 8th graders

The effect sizes of the overall differences between CEFR levels (A1 to B2) among the Finnish EFL learners were more prominent for the gymnasium students. A medium (but almost large) effect size was found for modifiers per noun phrase in the gymnasium group ($\eta 2$ = 0.13) and another medium effect size for left embeddedness ($\eta 2$ = 0.07) in the same group. The only effect size in the 8th-grade group related to overall significant differences across CEFR levels was small ($\eta 2$ = 0.05) and concerned left embeddedness.

Left embeddedness and the number of modifiers per noun phrase have yet to be investigated in SC research. Left embeddedness does not seem to have been studied concerning the CEFR much. The only CEFR-related study that included at least one was Green's (2012) investigation with EFL learners from various backgrounds. He found that the number of modifiers per noun significantly

separated between B2 and C1 and between C1 and C2 learners. However, McNamara et al. (2010) non-CEFR research found it to increase with proficiency, which is in line with the findings of the current dissertation for Sindhi-speaking EFL learners.

9.1.5 Phrasal density

Numerous new studies on SC have focused on communication at the phrasal level (Kyle, 2016). Coh-Metrix is suitable for analysing SC at this level since it offers a variety of phrasal-level indicators that quantify the density of various word types within a phrasal context (see Table 3.2). In Studies 1 and 2 of the current dissertations, several density indices separated the CEFR levels among Finnish and Sindhi learners. The values of some of the indices increased with skill level, while in other indices, the values decreased.

To summarise, in Study 1, comparing Sindhi and Finnish EFL learners at CEFR levels A1 to B1, the following findings were obtained (see Tables 8 and 12 in Supplementary Data and Figure 6 in Study 1):

- *noun phrase density*: significant overall decrease in the Finnish group but no changes in the Sindhi group; significant pairwise differences for A1 vs B1 and A2 vs B1 in the Finnish group; minimal variation among Sindhi learners in their use of noun phrases regardless of their CEFR level but considerable level-related variation among the Finns
- *Verb phrase density*: a significant overall increase in both L1 groups but significant pairwise only for A1 vs B1 in the Sindhi group; the Finns used more verb phrases than the Sindhi
- *adverbial phrase density*: a non-significant overall increase in the Finnish group but no change in the Sindhi group; minimal variation among Sindhi learners in their use of noun phrases regardless of their CEFR level but considerable level-related variation among the Finns
- *Preposition phrase density*: an overall increase in both L1 groups but significant only in the Sindhi group; significant pairwise for A1 vs A2 and A1 vs B1 in the Sindhi group; the Sindhi group used more preposition phrases than the Finns
- *negation phrase density*: significant overall decrease in both L1 groups; pairwise significant for A1 vs B1 (both L1 groups), and A1 vs A2 and A2 vs B1 in the Sindhi group; the Finns used somewhat more negation phrases than the Sindhi
- *gerund phrase density*: significant overall increase in both L1 groups; significant pairwise for A1 vs B1 (both L1 groups), and A1 vs A2 in the Sindhi group, and A2 vs B1 in the Finnish group; the Sindhi used somewhat more gerund phrases than the Finns

- *Infinitive phrase density*: a significant overall increase in both L1 groups; pairwise significant for A1 vs B1 in both groups, for A1 vs A2 for the Sindhi (marginally significant also for the Finns), marginally significant for A2 vs B1 for the Finns; the Finns at A2 and B1 used more infinitive phrases than their A2-B1 Sindhi peers

The most significant effect sizes for differences between CEFR levels were found for noun phrase density among the Finns ($\eta 2 = 0.102$), infinitive phrase density for the Finns ($\eta 2 = 0.081$), and negation phrase density for the Finns ($\eta 2 = 0.044$). For the Sindhi group, the effect sizes were somewhat smaller, and the largest ones were for negation phrase density ($\eta 2 = 0.043$), infinitive phrase density ($\eta 2 = 0.034$) and gerund phrase density ($\eta 2 = 0.017$) (see Table 8 in Supplementary data in Study 1).

In Study 2, focusing on the Finnish 8th grade and gymnasium students and covering CEFR levels A1 to B2, the following results were obtained (see Tables 12 and 13 and Figures 9 and 10 in Study 2):

- *Noun phrase density*: significant overall decrease in both age groups but significant pairwise only for the A2 vs B1 8th grade group (A2 vs B1 approached significance for the gymnasium group); 8th graders used somewhat more noun phrases than gymnasium students
- *verb phrase density*: significant overall increase among 8th graders but no discernible change among gymnasium students (pairwise A2 vs B1 approached significance among 8th graders); 8th graders used more verb phrases than gymnasium students
- *Adverbial phrase density*: a slight but significant overall increase in both age groups (pairwise A2 vs B1 approached significance among gymnasium group); gymnasium students used more adverbial phrases than 8th graders
- *Preposition phrase density*: a significant overall increase in both age groups but no significant pairwise differences; both age groups used roughly the same amount of preposition phrases
- *Negation phrase density*: decrease in both age groups but not significant overall and no significant pairwise differences; 8th graders used more negation phrases than gymnasium students
- *Gerund phrase density*: a significant overall increase in both age groups (pairwise A2 vs B1 approached significance among 8th graders); gymnasium students used more gerund phrases than 8th graders
- *Infinitive phrase density*: increase in both age groups but a significant overall increase only among 8th graders, who also had a significant pairwise increase for A1 vs A2 and A2 vs B1; both age groups used roughly the same amount of infinitive phrases

The size of the effect of the differences between CEFR levels was more prominent for the 8th graders. By far, the largest effect size for this group was for the infinitive phrase density ($\eta 2$ = 0.29). The next largest was for verb phrases ($\eta 2$ = 0.09), noun phrases ($\eta 2$ = 0.07) and preposition phrases ($\eta 2$ = 0.05). Infinitive phrase density was also the clearest CEFR level separator for the gymnasium group, even though the effect size was smaller than for the 8th graders ($\eta 2$ = 0.10). For them, noun, adverbial and preposition phrase densities all had the same effect size of 0.04.

The findings of this dissertation were like those of Kim (2004), who found the number of gerunds and infinitive phrases to increase as one progressed through the CEFR levels, even though Kim analysed phrasal unit proportions using sentences rather than units of 1,000 words. Also, in line with our findings, Green (2012) discovered that between B2 and C1, noun phrases become more frequent (denser). However, no change across A1-B1 in NP density could be found for the Sindhi learners in Study 1, and in Green's (2012) study of higher CEFR levels, a significant increase in NP density was found between B2 and C1. As an overall observation about the phrasal density indices of SC in the studies in this dissertation, it could be said that as proficiency in both language groups increased, particularly among the Finnish-speaking group, a shift away from noun and negation phrases occurred towards using more infinitive, gerund, verb, and adverbial phrases.

9.1.6 Similarity, variety, and simplicity

Similarity, diversity, and simplicity are three general SC measures in the Coh-Metrix. According to the findings in this dissertation, *syntactic similarity* and *syntactic simplicity* behaved as expected (if SC increases as proficiency increases). First, the similarity of adjacent sentences decreased with proficiency (adjacent sentences became syntactically more different). This was clearer among the Finnish learners. Second, syntactic simplicity also decreased with increasing proficiency (syntax became more complex); this trend was equally evident in both L1 groups. However, although the general pattern of changes in syntactic similarity and simplicity across CEFR levels was similar in both L1 groups, the absolute values were consistently higher for the Sindhi EFL learners, indicating they used more similar and simpler syntax compared to the Finnish EFL learners when the two groups' CEFR writing level was the same.

The third general SC index, called *minimal edit distance for parts of speech* in Coh-Metrix, refers to syntactic variation or diversity. It increased in the Finnish group, particularly between A1 and A2, but remained at the same level for the Sindhi EFL learners. For this index, too, the absolute values of the Sindhi learners were higher than those of the Finns, and even the A1 and A2 level Sindhis obtained values equal to those obtained only by the B1 level Finns.

To summarise, in Study 1, comparing Sindhi and Finnish EFL learners at CEFR levels A1 to B1, the following findings were obtained (see Tables 7 and 11 in Supplementary data and Figure 5 in Study 1):

- *Syntactic simplicity (z-score)*: decreasing overall trend but not significant in either L1 group; the Sindhi had higher simplicity values than their Finnish peers with the same CEFR levels
- *Syntactic simplicity (percentile)*: decreasing overall trend but not significant in either L1 group; the Sindhi had higher simplicity values than their Finnish peers
- *Sentence syntax similarity (adjacent sentences)*: overall decrease in both L1 groups that was marginally significant in the Finnish group; the Sindhi had higher similarity values than their Finnish peers
- *Minimal edit distance (parts of speech)*: a significant overall increase in the Finnish group (no trend in the Sindhi group); the significant pairwise difference for A1 vs B1 and marginally significant for A1 vs A2 in the Finnish group; the Sindhi had higher values than their Finnish peers

The effect size of the differences between CEFR levels was more prominent for the Finnish group. Two medium effect sizes were found for the Finnish group: $\eta 2 = 0.079$ for minimal edit distance and $\eta 2 = 0.023$ for sentence syntax similarity. None of the overall differences between CEFR levels was significant in the Sindhi group; therefore, the effect sizes were small.

In Study 2, focusing on Finnish EFL learners at CEFR levels A1 to B2, the following findings were obtained (see Tables 12 and 13 and Figures 7 and 8 in Study 2):

- *Syntactic simplicity (z-score)*: decreasing overall trend but not significant in either age group (although approached significance in the gymnasium group); the 8th graders had somewhat higher simplicity values than the gymnasium students with the same CEFR levels
- *Syntactic simplicity (percentile)*: decreasing overall trend but not significant in either age group (although approached significance in the gymnasium group); the 8th graders had somewhat higher simplicity values than the gymnasium students with the same CEFR levels
- *Sentence syntax similarity (adjacent sentences)*: an overall significant decrease in both age groups; the significant pairwise difference for A2 vs B1 in the gymnasium group; roughly similar similarity values in both groups

- *Minimal edit distance (parts of speech)*: a significant overall increase in the 8th-grade group and almost significant in the gymnasium group; the significant pairwise difference for A1 vs A2 and A2 vs B1 in the 8th-grade group and almost significant for B1 vs B2 in the gymnasium group; the gymnasium students had slightly higher or similar values than the 8th graders

The effect sizes of the differences between CEFR levels were larger for the 8th graders. A significant effect size was found for sentence syntax similarity across adjacent sentences in the 8th grade ($\eta 2$ = 0.22) and a medium effect size in the gymnasium ($\eta 2$ = 0.10). Minimal edit distance was the second largest overall separator of CEFR levels in this group of SC indices, with $\eta 2$ = 0.05 for the 8th graders and $\eta 2$ = 0.02 for the gymnasium group.

Green's (2012) research appears to be the first—and so far, the only, study to investigate these indices in the context of the CEFR; he also showed that as learners' skills increased from C1 to C2, their syntactic similarity decreased. Green's result suggests that the downward trend found in the current dissertation for lower CEFR levels may continue even at the highest end of the CEFR scale.

To conclude Section 9.1 on the differences in SC between CEFR levels, Table 9.1.6 summarises the most evident differences between the levels found in this dissertation. The table lists the ten largest **overall** effect sizes for specific SC indices across levels A1 to B1 in Study 1 and across levels A1 to B2 in Study 2 (effect sizes due to differences between pairs (e.g., adjacent) of CEFR levels are not included). For Study 1, the most prominent effect sizes are presented individually for the Pakistani and Finnish groups. In contrast, for Study 2, the effect sizes are presented for the two Finnish groups (the eighth-and twelfth graders).

Table 9.1.6 shows that the mean length of a sentence, the standard deviation of sentence length, the mean length of a T-unit, verb phrases per T-unit, infinitive density, dependent clauses per clause, and clauses per sentence have separated CEFR levels in Sindhi and Finnish groups (Study 1) significantly with the most significant effect sizes. Similarly, the number of words, number of clauses, number of complex T-units, number of the verb phrase, infinitive density, and sentence syntax similarity (adjacent sentences) have differentiated CEFR levels in 8th graders and Gymnasium groups (Study 2) significantly, with the largest effect sizes. Infinitive density is the only index which differentiated CEFR level learners significantly, with a large effect in both groups in Study 1 and the largest effect size in Study 2.

However, several indices, such as complex T-units per T-unit, noun phrase density, and minimal edit distance POS, have significantly differentiated the CEFR levels only in Finnish groups in Study 1. Similarly, a variety of unique indices, such as dependent clauses per T-unit, negation density, and left embeddedness, were revealed in Study 1 to separate the Sindhi groups' CEFR

levels significantly. Furthermore, in Study 2, the number of T-units, dependent clauses, sentence length, and sentence length (SD) significantly differentiated the CEFR levels in 8th-grade student groups. In study 2, modifiers per noun phrase, left embeddedness, the number of complex nominals, and clause length significantly differentiated the CEFR levels among gymnasium learners.

Table 9.1.6 Top 10 indices that significantly differentiated CEFR levels with the largest effect sizes.

Finns in Study 1	**Effect Size η2**	**Sindhis in Study 1**	**Effect Size η2**	**8th graders in Study 2**	**Effect Size η2**	**Gymn in Study 2**	**Effect Size η2**
Mean length of sentence	*0.248*	*Mean length of sentence*	*0.056*	*Sentence length*	.20	Modifiers per noun phrase	.13
Sentence length (S.D)	*0.186*	*Sentence length (S.D)*	*0.025*	*Sentence length (S.D)*	.24	Left embedded ness	.07
Mean length of T-unit	0.104	Mean length of T-unit	0.041	Number of Complex T-units	.37	Number of Complex T-units	.10
Verb phrases per T-unit	0.124	Verb phrases per T-unit	0.038	Number of Verb Phrase	.50	Number of Verb Phrases	.21
Infinitive density	0.081	Infinitive density	0.034	Infinitive density	.29	Infinitive density	.10
Dependent clauses per clause	0.097	Dependent clauses per clause	0.021	Sentence syntax similarity (adjacent sentences)	.22	Sentence syntax similarity (adjacent sentences)	.10
Clauses per sentence	0.084	Clauses per T-unit	0.029	Number of T-units	.35	Number of Complex nominals	.26
Complex T-units per T-unit	0.094	Dependent clauses per T-unit	0.025	Number of Dependent clauses	.45	Clause length	.110
Noun phrase density	0.102	Negation density	0.043	Number of words	.57	Number of words	.31
Minimal edit distance POS	0.079	Left embeddedness	0.019	Number of Clauses	.46	Number of Clauses	.11

Note: This table is primarily focused on examining the larger picture for each syntactic complexity index across the two studies. The goal was to identify any potential differences between the different groups in terms of their syntactic complexity. The focus was on the larger trends rather than individual cases, in order to see if there were any patterns that could be identified. Table 9.1.6 shows in *italics* the examples of mean sentence length and standard deviation for three groups in two studies.

9.2 Differences in syntactic complexity between specific CEFR levels

The findings of Studies 1 and 2 pertaining to changes in syntactic complexity across CEFR levels were summarised and compared to those of earlier studies on differences in SC across these competency levels. The preceding section focused on the various categories and individual SC indices. This section summarises the findings for pairs of adjacent CEFR levels (i.e., A1 vs. A2, A2 vs. B1, and B1 vs. B2) and two non-adjacent levels (A1 vs. B1) to supplement the discussion of the findings from Studies 1 and 2. The first section addresses the differences in syntactic complexity between the CEFR levels A1 and A2.

9.2.1 A1 vs. A2

This section summarises the findings for A1 and A2 learners. Beginner-level (e.g., A1 or A2) students have been identified as the most diverse students in each country, which suggests that the students within that level vary considerably. Second, the most significant number of SC differences discovered between A1 and A2 is noteworthy. In the first and second studies, around thirty-eight syntactic complexity indices separated the A1 from the A2 learners (see Table 9.2.1). Mean sentence length, sentence length (standard deviation), and verb phrases per T-unit were the best separators in all groups of learners across both studies. In addition, clauses per sentence were the clearest separator between A1 and A2 in the Finnish group of learners in the first study.

Table 9.2.1 Summary of SC indices that separated between CEFR levels A1 and A2 among Sindhi and Finnish learners of English (based on Studies 1 and 2)

A1 vs A2	**A1 vs A2**	**A1 vs A2**
Finns (1st study)	**Sindhis (1st study)**	**Finns (2ndStudy, 8th grade)**
Mean length of sentence **Sentence length (st. dev.)** *Verb phrases per T-unit* **Clauses per sentence** *T-units per sentence* *Minimal edit distance* *Modifiers per noun phrase* *Infinitive density*	**Mean length of sentence** **Sentence length (st. dev.)** **Mean length of T-Unit** **Verb phrases per T-unit** *Clauses per sentence* *Dependant clauses per clause* *Complex T-units per T-unit* **Clauses per T-unit** *Dependent clauses per T-unit* *Left embeddedness* **Infinitive density** **Negation density** *Preposition phrase density* *Gerund density*	**Mean sentence length** **Sentence length (st. dev.)** **Mean length of T-unit** Mean sentence length (st.dev.) *Clauses per sentence* *Clauses per T-unit* Minimal edit distance for parts of speech **Verb phrases per T-unit** Infinitive density *Dependent clauses per clause* *Dependent clauses per T-unit* *Coordinate phrases per T-unit* *Left embeddedness* *Minimal edit distance for PoS* *Verb phrase density* **Infinitive density**

Note: SC indices in **bold** font refer to indices that distinguished CEFR levels at $p < .001$; the indices in *italic* font refer to $.01 < p < 001$) and those with regular font to $.05 < p < .01$.

9.2.2 A2 vs. B1

This section summarises the findings for A2 vs. B1. Depending on the study, six to nine syntactic complexity indices separated A2 learners from B1 learners, totalling 25 times in studies 1 and 2 (see Table 9.2.2). The length measures were the indicators of syntactic complexity that differentiated the A2s most clearly from the B1s (e.g., sentence, T-unit, and clause length). In both studies, dependent clauses per clause/T-unit, complex nominals per T-unit, and sentence length (st. dev.) were the indices that distinguished A2 and B1 learners. In addition, A2 learners were separated from B1 learners by the number of modifiers per noun phrase, noun phrase, verb phrase, preposition, and negation phrase density.

Table 9.2.2 Summary of SC indices that separated between CEFR levels A2 and B1 among Sindhi and Finnish learners of English (based on Studies 1 and 2)

A2 vs B1	**A2 vs B1**	**A2 vs B1**
Finns (1st study)	**Sindhis (1st study)**	**Finns (2nd Study, 12th & 8th grade)**
Mean length of sentence **Mean length of sentence (st. dev.)** **Mean length of T-Unit** **Verb phrases per T-unit** *Dependant clauses per clause* *Complex T-units per T-unit* *Modifiers per noun phrase* **Noun phrase density** Infinitive density Gerund density	*Mean length of sentence* **Mean length of T-Unit** *Verb phrases per T-unit* *Dependant clauses per clause* *Complex nominals per T-unit* **Negation density**	*Mean length of sentence (st. dev.) (8)* *Mean length of sentence (st. dev.) (G)* **Minimal edit distance for parts of speech (8)** **Noun phrase density (8** *Noun phrase density (G)* *Adverbial phrase density (G)* **Infinitive density (8)** *Gerund density (8)* *Sentence syntax similarity (adjacent sentences) (G)*

Note: SC indices in **bold** font refer to indices that distinguished CEFR levels at $p < .001$; the indices in *italic* font refer to $.01 < p < 001$) and those with regular font to $.05 < p < .01$.
Note: 8 = grade eight; G = gymnasium

9.2.3 B1 vs. B2

This section highlights the results for B1 versus B2, based on the Finnish EFL learners in Study 2. In the Finnish setting, fourteen syntactic complexity indices significantly separated B1 and B2 level learners in the Gymnasium group (see Table 9.2.3). The most significant index was the number of words that separated B1 from B2 learners in the 8th and gymnasium groups (in fact, for the 8th graders, that was the only significant separator of B1 and B2 levels). For the gymnasium students, several SC indices distinguished between B1 and B2: the clearest ones were the number of complex T-units, number of complex nominals, complex nominals per clause, number of verb phrases, left embeddedness, and modifiers per noun phrase. The number of sentences, number of clauses, number of T-units, number of dependent clauses, mean length of the clause, complex nominals per T-unit, and minimal edit distance for PoS also separated the B1 from B2 learners among gymnasium students. Since there was only a small group of 8th-grade students with a B2 proficiency level, this lack of data was likely the main reason why only one SC index separated B1 and B2 levels in that group.

Table 9.2.3 Summary of SC indices that separated between CEFR levels B1 and B2 among Sindhi and Finnish L1 learners of English (based on Study 2)

B1 vs B2 based on 2nd study
Number of words (8)
Number of words (G)
Number of sentences (G)
Number of clauses (G)
Number of T-units (G)
Number of complex T-units (G)
Number of complex nominals (G)
Number of verb Phrase (G)
Number of dependent clauses (G)
Mean clause length (G)
Complex nominals per clause (G)
Complex nominals per T-unit (G)
Left embeddedness (G)
Modifiers per noun phrase (G)
Minimal edit distance for PoS (G)

Note: SC indices in **bold** font refer to indices that distinguished CEFR levels at $p < .001$; the indices in *italic* font refer to $.01 < p < 001$) and those with regular font to $.05 < p < .01$.

9.2.4 A1 vs. B1 learners based on 1st study

This section summarizes the findings for A1 vs. B1 learners, based on Study 1 (non-adjacent CEFR levels were not compared in Study 2). In Study 1, comparing A1 and B1 level learners, seventeen syntactic complexity indices distinguished these levels in the Finnish context and sixteen in the Pakistani context (see Table 9.2.4). The mean length of sentence, sentence length (St. Dev), mean length of T-unit, clauses per T-unit, verb phrases per T-unit, dependent clauses per clause, dependent clauses per T-unit, complex T-units per T-unit, complex nominals per T-unit and infinitive density were the best separators in both groups (Sindhis and Finns).

The clearest separators in the Finnish group of learners were the mean length of the clause, clauses per sentence, T-units per sentence, minimal edit distance negation density, and gerund density. The clearest separators in the Sindhi group of learners were clauses per sentence, gerund density, verb phrase density and preposition phrase density.

Table 9.2.4 Summary of SC indices that separated between the non-adjacent CEFR levels A1 and B1 among Sindhi and Finnish L1 learners of English based on Study 1

A1 vs B1 based on 1st study	
Finnish	**Sindhi**
Mean length of sentence	**Mean length of sentence**
Sentence length (st. dev)	**Sentence length (st. dev.)**
Mean length of T-Unit	**Mean length of T-Unit**
Mean length of clause	*Clauses per sentence*
Clauses per sentence	**Clauses per T-unit**
T-units per sentence	**Verb phrases per T-unit**
Clauses per T-unit	**Dependant clauses per clause**
Verb phrases per T-unit	**Dependent clauses per T-unit**
Dependant clauses per clause	**Complex T-units per T-unit**
Dependant clauses per T-unit	**Complex nominals per T-unit**
Complex T-units per T-unit	**Left embeddedness**
Complex nominals per T-unit	**Infinitive density**
Minimal edit distance	**Negation density**
Infinitive density	*Gerund density*
Noun phrase density	*Verb phrase density*
Negation density	*Preposition phrase density*
Gerund density	

The SC indices in **bold** font refer to indices that distinguished CEFR levels at $p < .001$; the indices in *italic* font refer to $.01 < p < 001$).

9.3 L1 and age-related effects on syntactic complexity in EFL writing

To complement the discussion of the findings of the current dissertation regarding the SC indices that distinguished between different CEFR levels (see Sections 9.1 and 9.2 above), the following sections focus on the other two areas covered in the dissertation: (1) the effect of the learners' first language on the syntactic complexity in their EFL writing, and (2) effect of the learners' age on the syntactic complexity in their EFL writing.

9.3.1 L1-related effects on syntactic complexity in EFL writing

The relationship between EFL learners' first language and syntactic complexity in their EFL writing was investigated only in Study 1, and, therefore, it is not necessary to synthesise the results concerning this topic in a similar way as those regarding the differences in SC across CEFR levels, which were investigated in two of the three articles that form this dissertation (i.e., Study 1 and Study 2). Hence, the principal results and their implications from Study 1 are briefly summarised here (for more details, see Study 1).

The main finding of Study 1 about the L1 effect was that the two groups differed considerably in terms of syntactic complexity in their EFL writing when their overall EFL writing level was comparable. The differences were more

explicit at the A2 and B1 levels, where most of the 28 SC indices investigated in the study differed between the two L1 groups. Level A1 was more comparable in terms of SC, but there, too, 12 of the 28 SC indices differed between the groups.

The most apparent differences between the two L1 groups relate to the length of the production units: The Sindhi-speaking EFL learners wrote longer clauses, T-units, and sentences than the Finnish-speaking EFL learners when their overall writing level was the same. The Sindhis used more coordination, and their sentences were more straightforward than those of the Finns, who used more subordinated sentences. The Sindhi's writing used phrasal-level features that tended to make their expressions longer, probably the main reason for their more extended production units. Such phrasal level features included higher left embeddedness, more complex noun phrases, and more nominals per clause and T-unit.

In Study 1 (Khushik & Huhta, 2020, p. 528), it was hypothesised that certain features of English used in Pakistan might contribute to the more extended expressions used by the Sindhi, such as the use of a "longer perfective aspect instead of the simple past (e.g., I had seen him yesterday instead of I saw him yesterday; Khan, 2012)." Further, in Study 1 (p. 528), it was hypothesised that the reason for greater left embeddedness in Sindhi speakers' EFL writing might be the fact that "because Sindhi is a subject-object-verb language, its L1 speakers may place more of the sentence elements before the verb when using a foreign language compared with the subject-verb–object languages such as Finnish" (see also Lashari and Soomro 2013).

Study 1 details a few similar studies related to the CEFR levels of EFL writing. The only previous study that used CEFR levels as the reference point for the EFL learners' proficiency and had several L1 groups was by Lu and Ai (2015). Their findings suggest differences in B2 and C1 level learners' SC in relation to their L1. Reviewing Lu and Ai's (2015) study, Study 1 (Khushik & Huhta, 2020, p. 211) reported that:

"For example, at B2, speakers of Japanese and Chinese differed from Tswana (from the Niger-Congo language family) speakers in the sentence and T-unit length, and particularly in clauses per T-unit, complex T-units per T-unit, and dependent clauses per clause/T-unit (as well as in clauses per sentence (pp. 23–24). "At the C1 level, Russian and German EFL learners differed in the length of production units and possibly in the proportional indices based on the clauses and T-units listed above, as well as in the number of clauses per sentence."

Two other studies have also compared SC in different L1-background EFL learners' writing. Bardovi-Harlig and Bofman's (1989) research investigated clauses per T-unit and included five L1 groups: Arabic, Chinese, Korean, Malay, and Spanish. The learners' overall English proficiency was the same based on TOEFL scores estimated to correspond to CEFR level B2. The study did not find significant differences in the SC index across the L1 groups. A different finding was obtained by Banerjee et al. (2007), whose study examined Chinese and

Spanish EFL learners with their International English Language Testing System (IELTS) writing performances. The study focused on the number of dependent clauses per clause and clauses per T-unit. Study 1 (Khushik & Huhta, 2020, p. 212) concluded that Banerjee et al.'s (2007) results "indicate that neither of the SC indices increased linearly across IELTS levels 3–8 in either of the groups." However, clauses per T-unit rose clearly between levels 4 and 5 (roughly A2/B1) among the L1 Spanish, whereas for the Chinese, it started increasing from level 5 onwards and was particularly pronounced between 7 and 8 (B2/C1).

Overall, the dissertation (notably Study 1 of it) contributed to our understanding of the linguistic generalisability of the CEFR levels across target languages (English in this case) and first languages (Sindhi and Finnish in this case). These results align with the two most extensive previous studies (Lu & Ai, 2015; Banerjee et al., 2007) investigating SC in EFL writing among different L1 groups. The two language groups in the current dissertation differ considerably as their first languages are unrelated, and these learners' sociolinguistic and educational contexts also differ. Therefore, the main result of Study 1 is that SC in EFL writing differs considerably depending on the learners' L1. This might not apply to EFL learners whose first languages are (closely) related, such as Italian and Spanish or Swedish and Norwegian. In Study 1, the SC indices that differed the least were sentence level similarity and simplicity, subordination, and specific phrasal indices. Whether these SC indices are more generalisable across languages than others remain to be investigated in future research.

9.3.2 Age-related effects on syntactic complexity in EFL writing

The relation between age and syntactic complexity in EFL was examined in the dissertation in Study 2 and a more focused manner in Study 3. Study 2 investigated the differences across CEFR levels by examining two age groups in Finland: 8th graders (primarily 14-year-olds in the comprehensive school) and 12th graders (mostly 17-year-olds in the public secondary school or gymnasium). The two age groups were not directly compared with statistical analyses. However, looking at the differences in the results of these two age groups can shed some light on possible age-related differences among Finnish EFL learners. In a limited way, the findings can be compared with those obtained for the Finnish learners in Study 3, which focused explicitly on systematically comparing learners of different ages. Study 3 included Pakistani and Finnish EFL learners and investigated if learners from those countries used equally complex syntax in their EFL writing when their overall CEFR writing level was the same but when their ages differed. Like Study 2, this study compared 8th graders and 12th graders. However, in contrast to Study 2, the data comprised writing performances on one task only that was the same in both age groups, whereas, in Study 2, the two Finnish age groups wrote three tasks, and their CEFR level was determined by considering their performance on all the tasks.

Tables 9.3.1 and 9.3.2 summarise the key age-related findings from Studies 2 and 3. Table 9.3.1 displays the SC indices that most clearly separated CEFR levels in terms of their effect sizes in the two age groups in Study 2. Table 9.3.2

displays whether SC indices showed statistically significant differences for the identical proficiency level learners whose age differed, separately for Pakistan and Finland.

The overall picture from Study 2, which focused on Finnish 8th and 12th graders (Table 9.3.1), was that the overall differences across CEFR levels were larger among the younger age group. Altogether five large, that is, $\eta 2 = 0.14$ or higher (see Richardson, 2011), effect sizes were found for the 8th graders: infinitive phrase density, the standard deviation of sentence length, syntax similarity across adjacent sentences, mean sentence length and mean T-unit length. No large effect sizes were found for the 12th graders, but in the case of six SC indices, at least medium (0.06 – 0.13) effect sizes were obtained for modifiers per noun phrase and mean clause length (for example, see Table 9.3.1). In addition to these effect size differences, there were more statistically significant overall differences among the younger group than in the older group (see Study 2 for details).

Regarding similarities and differences between the two-age group based on Study 2 results, some tentative trends can be seen, which may provide ideas and hypotheses for further research in the future. The context of Study 2 was a country with a relatively homogeneous language education culture that both age groups had gone through, even though individual learners' free-time use (and, thus, learning) of English can vary substantially. Furthermore, the 8th graders represent the whole age group. In contrast, the 12th graders represent a more selective section of their age group since gymnasia tend to attract more academically oriented students as it is the main gateway to university and polytechnic level of education.

The similarities seen in Table 9.3.1 include the finding that in both age groups, infinitive phrase density and syntax similarity across adjacent sentences were among the clearest separators among the learners at different CEFR levels. In addition, mean sentence length, the standard deviation of sentence length, and noun phrase density were among the suitable separators. The most noticeable difference between the age groups was the difference in effect sizes mentioned above. Syntactic complexity was a more important factor among the younger learners in distinguishing their EFL proficiency levels than among the somewhat older EFL learners, at least in Finland. Regarding individual SC indices, the mean length of the longer production units (sentences, T-units) was an excellent separator among the younger learners. In contrast, mean clause length was among the best two separators in the older age group (mean sentence length separated in that group, too, but with only a small effect size).

Some caveats are in order regarding comparing the findings for the two age groups of Finnish EFL learners in Study 2. The first is that the differences between the two groups are not purely related to their age but also to differences in the length of their formal education in English and any exposure to English in their free time (this point is also made elsewhere in this dissertation). Therefore, even if the learners' CEFR writing level was rated to be comparable, the two groups' texts (at any CEFR level) were not identical when it came to the various linguistic

features in them. Study 2 showed that some syntactic complexity features differed in the two groups, at least in that their importance for distinguishing between CEFR levels varied. However, since the quality of writing depends on several factors, not just syntax and its complexity, the raters using the can-do/communication oriented CEFR scales placed the learners' scripts at the various proficiency levels by considering a wide range of criteria (see Appendix 2). Syntactic complexity is likely to have played some role in the raters' minds when they decided on the CEFR level of the scripts. However, the influence of syntax must have been indirect since the CEFR scales used for rating contained very few expressions that could be seen to refer to syntactic patterns (see Study 2 for a discussion about the raters and the potential role that the linguistic aspects of learners' texts may have played in the ratings).

The second caveat relates to the different distribution of the CEFR levels in the two age groups in Study 2, which may have affected the results of the statistical analyses of the data (see Table 3 in Study 2). The CEFR levels in the 8th graders' data ranged from A1 to B2, but there were only eight B2-level scripts. In contrast, the 12th-grade data only contained scripts from A2 to B2, and had more scripts at the B1 level than the other two levels; this might have contributed to the somewhat smaller number of significant overall differences across CEFR levels and the smaller effect sizes found for the 12th graders.

Table 9.3.1 Ten largest overall effect sizes for the SC indices across the CEFR levels for the 8th and 12th graders in Finland in Study 2

8th grade		12th grade	
SC index	effect size ($\eta 2$)	SC index	effect size ($\eta 2$)
infinitive phrase density	0.29	modifiers per noun phrase	0.13
standard deviation of sentence length	0,24	mean clause length	0.11
syntax similarity across adjacent sentences	0.22	infinitive phrase density	0.10
mean sentence length	0.20	syntax similarity across adjacent sentences	0.10
mean T-unit length	0.14	left embeddedness	0.07
verb phrases per T-unit	0.10	complex nominals per clause	0.07
verb phrase density	0.09	standard deviation of sentence length	0.05
clauses per sentence	0.09	mean sentence length	0.04
clauses per T-unit	0.08	noun phrase density & adverbial phrase density & preposition phrase density	0.04
noun phrase density	0.07		

Study 3 concerned age-related differences, too, but the picture emerging from it differs from the picture from Study 2 because the designs and analyses in the two studies were different. Study 3 used only performances on one task that was the same in both age groups (8th and 12th graders) in both countries and L1 groups (Sindhis in Pakistan and Finns in Finland) and compared learners with a different age but the same CEFR level to find out if syntactic complexity was the same in the two age groups when their overall writing level was comparable. Thus, possible differences across CEFR levels were not investigated in Study 3; instead, A1-level 8th graders were compared with A1-level 12th graders, A2s with A2s, and B1s with B1s.

As Table 9.3.2 shows, the effect of age was much more evident among the Sindhis than among the Finns. The mean length of sentence, mean length of T-unit, dependent clauses per T-unit, coordinate phrases per T-unit, syntactic simplicity (z-score), syntactic simplicity (percentile), and negation density were found to be significantly different in the two-age groups at each of the CEFR levels A1, A2, and B1 in the Sindhi group. In addition, dependent clauses per clause, complex nominals per T-unit, and left embeddedness significantly differed at the A2 level in the Sindhi group. Interestingly, almost all indices significantly differed at the A1 level in the Sindhi group. The significant indices at the B1 level in the Sindhi group were clauses per sentence, verb phrases per T-unit, and clauses per T-unit. The only SC indices showing age-related differences for the Finnish group were the mean length of clauses (for A2 and B1 levels) and clauses per sentence (for B1). Study 3 supports Study 1 because the two L1 groups differed considerably. Whereas Study 1 showed that L1, in general, can have a significant effect on the syntactic complexity in EFL writing when the different L1 learners' overall writing proficiency level in English as a foreign language is the same, Study 3 complemented that picture by showing that these different L1 groups also differ in that their age, too, seems to impact their syntactic complexity. Age matters more in Pakistan than in Finland when syntactic complexity in EFL writing is concerned.

Table 9.3.2 Significant differences in SC indices between 8th and 12th graders whose CEFR level was the same (in Study 3)

SC index	A1	A2		B1	
	Sindhi	Sindhi	Finnish	Sindhi	Finnish
Mean length of Sentence length	X	X		X	
Sentence length (S.D)	X			X	
Mean Length of T-Unit	X	X		X	
Mean Length of Clause	X		X		X
Clauses per sentence	X			X	X
Verb Phrases per T-unit	X			X	
Clauses per T-unit	X			X	
Dependant clauses per clause	X	X			
Dependant clauses per T-unit	X	X		X	
T-unit per sentence	X				
Complex T-units per t-unit	X				
Coordinate Phrases per T-unit	X	X		X	
Complex nominals per T-unit		X			
Complex nominals per clause	X				
Syntactic simplicity, z-score	X	X		X	
Syntactic simplicity, percentile	X	X		X	
Left embeddedness	X	X			
Sentence syntax similarity, adjacent sentences, mean	X				
Sentence syntax similarity, all combinations, across	X				
Adverbial phrase density	X				
Negation density	X	X		X	

X in Table 9.3.2 denotes a statistically significant difference between the 8th and 12th graders at a particular CEFR level and L1 group

Given the different purposes and analyses used in Study 2 and Study 3, no clear conclusions can be drawn about the findings' similarities, even if age was a factor investigated in both, at least for Finnish EFL learners. It can, however, be noted that the only two SC indices that showed age-related differences in Study 3 for the Finnish learners, namely the mean length of clause and clauses per sentence, also turned out to be among the clear overall separators of proficiency levels for the Finnish learners. However, mean clause length was such a separator only in the older group and clauses per sentence only in the younger group.

9.3.3 Variation in the values of SC indices across studies

This section summarises the findings of the current dissertation and previous CEFR-related research on syntactic complexity by examining the *values* obtained *for specific SC indices*. The focus of the dissertation has been on the differences in SC between CEFR levels (Studies 1 and 2), between different L1 groups (Study 1), and between different age groups (Study 3, and partially

also Study 2). This section focuses on the *characteristics of each CEFR level in terms of syntactic complexity* (in EFL writing) *in numerical terms*. This section, thus, expands the comparison of the SC values obtained in Study 1 and those reported by Banerjee et al. (2007) in Table S17 in Supplementary Materials for Study 1. Such level-by-level analysis of the linguistic features (e.g., values of SC indices) is another angle in investigating the linguistic basis of the CEFR levels. It complements research on the linguistic differences between the CEFR levels (see Khushik & Huhta, 2020, p. 529). This analysis resembles, thus, the work done, for example, in the Profile English project (see Green, 2012; Hawkins & Filipovic, 2012), which is based on the Cambridge English examination responses that provide information about the linguistic features (e.g., words, structures) that are typical of EFL learners writing at CEFR levels. However, here the aim is to characterise CEFR levels with numerical values rather than concrete linguistic samples.

This investigation of SC values is not meant to be exhaustive but to illustrate how similar or different values for, say, the mean length of sentences or clauses, have been obtained for CEFR levels, say, B1, in different studies of EFL writing. Table 9.3.3 presents these values. The differences between different studies on SC are due to several factors, such as the L1 and age (or educational experience) of the learners, but also the data (one or more writing tasks, text type, genre) from which the SC indices were calculated, and how the learner performances (scripts) were placed on the CEFR levels.

In Table 9.3.3, we can see, for example, that Polat (2020) reported a sentence length of roughly eight words for B1 proficiency level learners, whereas Martinez (2018) reported over nineteen-word sentence lengths, and Qi (2014) reported high values of twenty-two words (see also Appendix 1, Table 2). When studying age effects, it is possible to see a continuous relationship between age and syntax. Time spent learning a language increases proficiency, as the SC values tend to increase from A1 to A2 to B1. A similar tendency can also be seen within a particular CEFR level. As established in this dissertation and prior research, adult learners' mean values for a range of syntactic complexity indices are often higher than for young learners.

Previous research has demonstrated that language learning is affected by many factors, including the type of interaction, motivation, time spent on learning and utilising what was learnt, type of classroom instruction, attitudes toward social integration, linguistic ability, working memory, and attention management, as well as life span ageing (see, e.g., Charisee, 2014, pp. 5-33). The data regarding the EFL writing of young and adult learners in the current and previous research (see Table 9.3.3) revealed various tendencies and patterns. Syntactic differences exist between learners of various ages and first languages, even when their writing proficiency is the same.

Specific trends emerged from the CEFR-related studies on syntactic complexity and the current investigation (see Table 9.3.3 and Appendix 1, Table 2) when three kinds of data sets are analysed: (1) only adult learners, (2) only young learners, and (3) mixed learners. A clear trend is that only a few studies

have been conducted on A1 learners (or on C1 or C2 learners, for that matter). Another trend appeared when the four groups of adults, mixed learners, and young learners were compared: the mean values for almost all syntactic complexity indices in the writings of young learners are lower than those for adults and mixed learners.

Table 9.3.3 Comparison of the values for different SC indices obtained in the current dissertation (Study 1, 2 and 3) to those obtained in previous research

SC features	Young learners in Study 2 & 3 (14–15-year-old)	Young learners in previous studies	Young adult learners in Study 2 & 3 (17–18-year-old)	Adult learners in previous studies	Mixed (young & young adults) learners in Study 1
B1 CEFR level					
Mean sentence length	15.13 (Finns), 12.61 (Sindhis) 12.66 (8th Finns)	19.93 (M)	13.61 (Finns), 16.36 (Sindhis) 14.24 (12th Finns)	10.8 (H & F), 7.61 (P) 16.49 (Qi)	13.80 (Finns) 14.78 (Sindhis)
Mean T-unit length	14.38 (Finns), 14.44 (Sindhis) 11.42 (8th Finns)	11.62 (G) 10.22(V)	14.53 (Finns), 17.28 (Sindhis) 12.73 (12th Finns)	7.10 (P)	14.62 (Finns) 16.21 (Sindhis)
Mean clause length	7.13 (Finns), 10.07 (Sindhis) 6.59 (8th Finns)	6.37 (G)	8.49 (Finns), 9.87 (Sindhis) 7.91 (12th Finns)	4.91 (P)	8.33 (Finns) 9.96 (Sindhis)
Clauses per T-unit	2.03 (Finns), 1.49 (Sindhis) 1.76 (8th Finns)	1.83 (G)	1.79 (Finns), 1.92 (Sindhis) 1.66 (12th Finns)	0.87 (P) 1.71 (Chi5; B) 2.59 (Spa5; B)	1.83 (Finns) 1.66 (Sindhis)
Dependant clauses per clause	0.45 (Finns) 0.29(Sindhis) 0.32 (8th Finns)	0.46 (M)	0.43 (Finns) 0.33 (Sindhis) 0.37 (12th Finns)	0.23 (P)	0.43 (Finns) 0.32 (Sindhis)
Modifiers per noun phrase	0.51 (Finns) 0.64 (Sindhis) 0.43 (8th Finns)		0.55 (Finns) 0.62 (Sindhis) 0.56 (12th Finns)	0.93 (GR)	0.55 (Finns) 0.62 (Sindhis)

(continues)

Table 9.3.3 continues

SC features	Young learners in Study 2 & 3 (14–15-year-old)	Young learners in previous studies	Young adult learners in Study 2 & 3 (17–18-year-old)	Adult learners in previous studies	Mixed (young & young adults) learners in Study 1
Sentence syntax similarity, adjacent sentences	0.09 (Finns) 0.13 (Sindhis) 0.10 (8th Finns)		0.08 (Finns) 0.11 (Sindhis) 0.10 (12th Finns)	0.11 (GR)	0.07 (Finns) 0.11 (Sindhis)
A2 CEFR level					
Mean sentence length	10.77 (Finns), 12.34 (Sindhis) 11.49 (8th Finns)	12.16 (M)	12.24 (Finns), 14.36 (Sindhis) 13.06 (12th Finns)	15.50 (Qi) 7.9 (H & F), 7.13 (P)	11.58 (Finns) 13.56 (Sindhis)
Mean T-unit length	11.73 (Finns), 13.41 (Sindhis) 11.08 (8th Finns)	7.10 (G) 8.18 (V)	12.68 (Finns), 15.95 (Sindhis) 12.01 (12th Finns)	6.90 (P)	12.45 (Finns) 14.96 (Sindhis)
Mean clause length	7.02 (Finns), 9.29 (Sindhis) 6.38 (8th Finns)	5.54 (G)	8.45 (Finns), 9.97 (Sindhis) 7.42 (12th Finns)	4.60 (P)	7.60 (Finns) 9.74 Sindhis)
Clauses per T-unit	1.63 (Finns), 1.59 (Sindhis) 1.77 (8th Finns)	1.30 (G)	1.51 (Finns), 1.69 (Sindhis) 1.67 (12th Finns)	0.85 (P) 1.54 (Chi4; B) 1.77 (Spa4; B) 1.44 (Chi3; B)	1.71 (Finns) 1.58 (Sindhis)
Dependant clauses per clause	0.34 (Finns) 0.24 (Sindhis) 0.30 (8th Finns)	0.30 (M)	0.33 (Finns) 0.30 (Sindhis) 0.34 (12th Finns)	0.20 (P)	0.35 (Finns) 0.28 (Sindhis)
Modifiers per noun phrase	0.46 (Finns) 0.57 (Sindhis) 0.42 (8th Finns)		0.50 (Finns) 0.60 (Sindhis) 0.51 (12th Finns)	0.62 (GR)	0.46 (Finns) 0.61 (Sindhis)
Sentence syntax similarity, adjacent sentences	0.09 (Finns) 0.13 (Sindhis) 0.11 (8th Finns)		0.11 (Finns) 0.11 (Sindhis) 0.13 (12th Finns)	0.12 (GR)	0.08 (Finns) 0.12 (Sindhis)

continues

Table 9.3.3 continues

SC features	Young learners in Study 2 & 3 (14–15-year-old)	Young learners in previous studies	Young adult learners in Study 2 & 3 (17–18-year-old)	Adult learners in previous studies	Mixed (young & young adults) learners in Study 1
A1 CEFR level					
Mean sentence length	8.04 (Finns) 10.30 (Sindhis) 8.58 (8th Finns)		13.99 (Sindhis)		
Mean T-unit length	9.73 (Finns) 12.57 (Sindhis) 8.53 (8th Finns)	7.43 (V)	14.81 (Sindhis)		11.10 (Finns) 13.84 (Sindhis)
Mean clause length	7.21 (Finns) 10.76 (Sindhis) 5.97 (8th Finns)		9.74 (Sindhis)		
Clauses per T-unit	1.41 (Finns) 1.28 (Sindhis) 1.44 (8th Finns)		1.59 (Sindhis)		
Dependant clauses per clause	0.24 (Finns) 0.22 (Sindhis) 0.24 (8th Finns)		0.27 (Sindhis)		
Modifiers per noun phrase	0.55 (Finns) 0.63 (Sindhis) 0.45 (8th Finns)		0.62 (Sindhis)		
Sentence syntax similarity, adjacent sentences	0.12 (Finns) 0.15 (Sindhis) 0.14 (8th Finns)		0.11 (Sindhis)		

Abbreviations used in the table:
Alexopoulou et al. (2017) = (A); L1 varied, probably mostly adult learners.
Banerjee et al. (2007) = (B) (Chi3 / Chi4 / Chi5 = Chinese EFL learners at IELTS band 3, 4 or 5; Spa4 / Spa5 = Spanish EFL learners at IELTS band 4 or 5), probably mostly adult learners.
Green (2012) = (GR); L1 varied, age varied but probably mostly (young) adults.
Gyllstad et al., (2014) = (G); Swedish EFL learners, 10-19 years old.
Hawkins and Filipović (2012) = (H & F); L1 varied, age varied but probably mostly (young) adults.
Martinez (2018) = (M); L1 Spanish EFL learners, 12-16 years old.
Polat et al., (2020) = (P); L1 Turkish EFL learners, 18-22 years old.
Qi (2014) = L1 varied, mostly adult learners.

A more detailed analysis of the studies shown in Table 9.3.3 allows us to see if some SC indices show more vs less variation when the learners' EFL proficiency/writing level is the same. Such an analysis may shed light on which SC indices in EFL writing are more vs less affected by such factors as learners' age, educational level or first language. Potentially, the analysis can also help identify studies whose results differ considerably from the others, which may mean that a closer look at those particular studies is needed to find out the reason why their results differ so much from those of the others (e.g., if it is because of the unique characteristics of the learner group or because the estimation of the learners' and texts' CEFR level was inaccurate).

However, previous CEFR-related SC research has been limited, particularly regarding the number of different SC indices examined in those studies. Therefore, most of the data on the values of selected SC indices presented in Table 9.3.3 come from the three studies that form the core of the current dissertation. A quick look at the values for the various SC indices at CEFR levels in Table 9.3.3 reveals that most of the values that differ from the most typical values obtained in research come from studies carried out outside the current dissertation. Thus, even if significant and apparent differences were obtained in the current dissertation for the two L1 and age groups, these differences were often more minor than those found in the other studies. This is an exciting finding and is not easy to explain. One possible reason might be that the CEFR levels in the current dissertation and the other studies have differed. The current dissertation's rating procedures (scales, training) were the same across the two contexts (Pakistan and Finland). The analyses showed that the raters in both countries had applied the rating scale sufficiently similarly. As was discussed earlier in this dissertation (and, e.g., in Study 1), operationalising the CEFR levels in the same way across different studies cannot be taken for granted. Therefore, it is possible that the CEFR levels in some of the other studies included in Table 9.3.3 were not the same as those in the current dissertation. In addition, the tasks completed by the learners can affect the values of SC indices, as can the learners' age (or educational level), at least to some extent, which is illustrated by comparing the values obtained for the Finnish EFL learners in Study 1, 2 and 3 (see, e.g., their values for the mean length of sentence at B1 level in Table 9.3.3).

The following observations about the typical values for the selected SC indices in Table 9.3.3 can be made:

- *The mean length of the sentence*. In most learner groups (at least for those involved in the current dissertation), the mean number of words per sentence at the B1 level has ranged from 13 to 16. Martinez (2018) and Polat et al. (2020) are outliers in this respect, with a mean length of 20 in the former study and 8 in the latter. Hawkins and Filipović (2012) were also somewhat different from most other studies with their 11-word sentence length. At level A2, the typical mean sentence lengths seem to be between 11 and 14

words. Again Polat et al. (2020) differs from the others with a low mean of only seven words per sentence. Hawkins and Filipović (2012), too, found only about eight words per sentence at the A2 level in the data they studied. Alexopoulou et al. (2017), in contrast, found somewhat longer mean sentences of about 15 words in their study. Very few studies exist for the A1 level (only those that are part of this dissertation, to my knowledge); mean sentence length ranged from 9 to 14 words, and there appear to be significant differences due to L1 or the age of the learners.

- *Mean T-unit length*. The T-unit length was usually between ca. 12 and 17 words at the B1 level. The only clear outlier was seven, found by Polat et al. (2020). The range was roughly the same for the A2 level, but three previous studies found shorter T-units of 7 or 8 words (Green, 2012; Verspoor et al., 2012; Polat et al., 2020). T-unit length also varied considerably among A1 level learners, typically between 9 and 14 words.
- *Mean clause length*. Typical clauses at the B1 level were from 7 to 10 words in length, with again Polat et al. (2020) as the most different with five words and Green (2012) with six words. For the A2 level, the typical range was almost the same (6 – 10 words), with Polat et al. and Green being somewhat different (5 and 6 words, respectively). Variation in clause length at A1 was also quite considerable in the available studies (from 6 to 11 words).
- *Clauses per T-unit.* Typical values for this SC index ranged from about 1.5 to 2.0 at B1. The outliers were Polat et al. (2020) with 0.87 and one of the groups in Banerjee et al. (2007) study with 2.59. For the A2 level, variation was minor in the 1.5 to 1.8, except for Polat et al. (0.85) and Greene (1.30). At A1, the values were somewhat lower: between 1.3 and 1.6.
- *Dependent clauses per clause*. At the B1 level, most values were between 0.3 and 0.4, with Polat et al. (2020) being somewhat lower at 0.23. For the A2 level, most values were lower between ca. 0.25 and 0.35, with Polat et al. being lower at 0.20. For A1, the values were slightly smaller, and there was less variation, as the values ranged from 0.22 to 0.27.
- *Modifiers per noun phrase*. Most values ranged from ca. 0.4 to 0.6, with Green (2012) being an outlier with 0.93 at level B1. At level A2, a similar range from 0.4 to 0.6 was found in all the studies, as was the case at A1.
- *Sentence syntax similarity for adjacent sentences.* This SC index ranged from 0.07 to 0.11 at level B1 and from 0.08 to 0.13 at A2. At A1, it was 0.11–0.15.

Such comparison across studies, as was done above, illustrates how syntactic complexity in EFL at the different CEFR levels can be described concerning the values in the SC indices. As more research is carried out on SC in EFL writing across different contexts, L1s and age groups, it may be possible to characterise each CEFR level in terms of typical values each SC index takes. Some indices are elementary to understand; for example, the mean lengths of clauses, T-units and sentences are transparent and readily comprehensible, but others, such as sentence syntax similarity, are much more difficult.

The above analysis of the values obtained in CEFR-related research does not allow us to say which SC indices are more generalisable across different contexts, tasks, L1s and age groups. They may all be affected by the various factors that can influence the syntactic complexity of learner texts. The analysis, however, may shed some light on which SC indices show more apparent changes (either increase or decrease) across CEFR levels, particularly from A2 to B1, which are the CEFR levels most often studied in the current and previous research. The changes from A2 to B1 were captured best by mean sentence length (but less by mean T-unit and clause lengths), clause per T-unit, dependent clauses per clause, and sentence syntax similarity (but not by modifiers per noun phrase). However, Table 9.3.3 includes only those SC indices most often studied in CEFR-related research. These conclusions can be biased because most of the groups in the table are those investigated in the three studies in the current dissertation.

9.4 Implications

This transdisciplinary dissertation draws on the theoretical concepts, approaches, methods, and tools developed in second language acquisition and language testing research, corpus linguistics, and language technology research. The present study sheds light on the linguistic characteristics of the CEFR levels by focusing on how these levels differ in terms of syntactic complexity. The study is purely quantitative, making it more like basic research than applied research in that its results are not directly applicable as qualitative descriptions of syntactic complexity across the CEFR scale. However, Section 9.3.2 above shows how particular CEFR levels could be described concerning quantitative indices of SC. Turning such numerical information into verbal descriptions is not straightforward and will require further effort. The present study has shown that however quantitative information about the differences in SC between CEFR levels (or typical SC values at different CEFR levels) is used, it is essential to consider the L1 background of the learners and probably also their age since many SC indices are likely to be affected by such learner characteristics.

One application of the research results, such as the one carried out in this dissertation, is in using them as predictors of the (CEFR) level of the texts written by English language learners. Information about typical numerical values at different CEFR or other levels could be part of an automated assessment system that aims to produce a grade or feedback to SFL English learners. Such predictive

systems would calculate a wide range of linguistic and textual features, not syntax, to increase their accuracy. Perhaps the best-known system that predicts the CEFR level of texts written by EFL learners is the Text Inspector (https://textinspector.com/), the development of which was based on the English Profile project and data. Some recent studies have also reported on the development and evaluation of such systems; for example, Arnold et al. (2018) and Gaillat et al. (2021) used a variety of lexical, syntactic, cohesive and accuracy features to predict the CEFR level of English language learners.

Furthermore, the CEFR is not commonly used in Pakistan, at least not yet. Thus, the research can have practical implications for English language development, instructions, and curricula, notably in the Pakistani educational institutions participating in the data collection, rater training, and rating of learners' scripts for this dissertation since those institutions and individuals in them obtained first-hand experience about the meaning of the CEFR levels.

9.5 Evaluation of the study: Strengths and Limitations and future research directions

The strong points of the current study include the relatively large number of learners involved, the range of syntactic complexity indices covered, and careful data collection and analysis that used procedures developed in language testing research.

The amount of data and the number of SC indices allowed the researcher to obtain a richer picture of how SC distinguishes between the CEFR levels; this also made it possible to compare the findings with many previous studies that each had typically covered only a few SC indices. The data also increased the statistical power of the analyses in many comparisons between the CEFR levels or L1 groups (although sometimes too few cases were left in particular groups, such as B2 level learners in the 8th graders' group in Study 2).

Another strength of the study was its application of several procedures borrowed from language testing research to ensure the validity of the data. This included the data gathering instruments, i.e., the writing tasks that came from previous research studies (the CEFLING and TOPLING projects) and were, thus, proven to work to assess language learners' writing ability. This also included rater training procedures that relied on practice ratings, consensus-seeking discussions, and benchmark samples. Finally, this included using the multifaceted Rasch rating analyses to ensure sufficient quality was achieved in estimating learners' EFL writing ability and the placement of the scripts on the CEFR levels.

While this research examined EFL writing at CEFR levels A1, A2, B1, and B2, the top two CEFR levels (and B2 among the Pakistani EFL learners) were not covered. Therefore, to get a more comprehensive picture of the whole range of proficiency, research needs to be broadened to encompass the CEFR levels C1

and C2, which have been studied less than the other levels. However, A1 level learners should also be investigated more, as apart from the current dissertation, SC at that level has received very little attention in previous CEFR-related research.

Although this study focused on writing abilities, the research approach may address other abilities (mainly speaking) in the future.

This study also paves the way for future research into whether there are variations when learners come from various situations and have a variety of L1 backgrounds from various nations. Furthermore, L1 and age-related differences and similarities could be investigated with other linguistic and textual features to widen the range of features from the syntactic characteristics examined in this thesis.

The current research did not explore differences that the writing task or the genre of the writing task might have on learners' syntactic complexity. Previous research suggests that the task can influence the language produced by the learner. The present research also found that the learners' SC varied depending on how their SC was calculated (e.g., based on only one writing task as in studies 1 and 3, or three writing tasks as in study 2). Thus, more research on the effect of the task(s) on syntactic complexity is also needed. Since the Pakistani learners completed several writing tasks, only one of which was used in the present study, those performances could be used in future research.

Research conducted by Kyle (2016) as well as Kyle and Crossley (2017 & 2018) suggest that fine-grained phrasal level indices of syntactic sophistication and complexity may be more accurate in predicting English language learners' writing skills (and thus, more accurate in separating CEFR levels) than traditional indices of absolute syntactic complexity. Kyle (2016) and Kyle and Crossley (2017 & 2018) suggest that looking at the complexity of specific syntactic constructions (such as subordinate clauses or relative clauses) may be a better indicator of writing proficiency than traditional text length metrics. However, according to the current study, the length of a text is a good way to measure someone's proficiency. This is one of the traditional or holistic ways to measure skill. The study found that the amount of text can be a reliable way to determine someone's ability. This is one of the classic or all-encompassing standards of measuring skill. Nevertheless, the findings of the current dissertation emphasize the importance of analyzing overall syntax and structure. This is in order to ensure that the level of writing proficiency is accurately assessed.

During my research, I faced a challenge regarding the verification of tool accuracy, specifically pertaining to their parsing abilities. Despite my efforts, I could not locate any studies that examine the accuracy of Coh-matrix or L2SCA. However, my primary objective was to assess the overall utility of these tools rather than their precision. Ultimately, based on the descriptions regarding the accuracy of the tools by Lu (2014) and McNamara et al. (2014), I found that both Coh-matrix and L2SCA were valuable resources in analyzing text complexity. Second, a reanalysis of the data collected for this dissertation using an automated analysis tool such as TAASSC (Tool for the Automatic Analysis of Syntactic

Sophistication and Complexity; Kyle, 2016) could provide useful insight into how CEFR levels differ in terms of phrasal complexity. Automated tools such as TAASSC are designed to analyze language data quickly and accurately. By using this tool to analyze the data, it would be possible to compare the complexity of phrasal structures in different CEFR levels and gain a better understanding of how they differ.

A further way to re-analyze the data could be done by using linear mixed-effect models in the future since they can simultaneously model the effect of CEFR level, L1 background, age (grade level), and task on SC. Linear mixed-effect models allow for the consideration of different variables and their interactions with each other, as well as the ability to account for individual differences that may influence the overall outcome. This could provide valuable insights into the effect of the different factors investigated in the current study. This is like building a complex jigsaw puzzle, taking into account the different shapes and colors of the pieces, as well as the individual characteristics of the pieces themselves. The resulting image would provide a more nuanced understanding of the overall picture than would have been possible with just one piece.

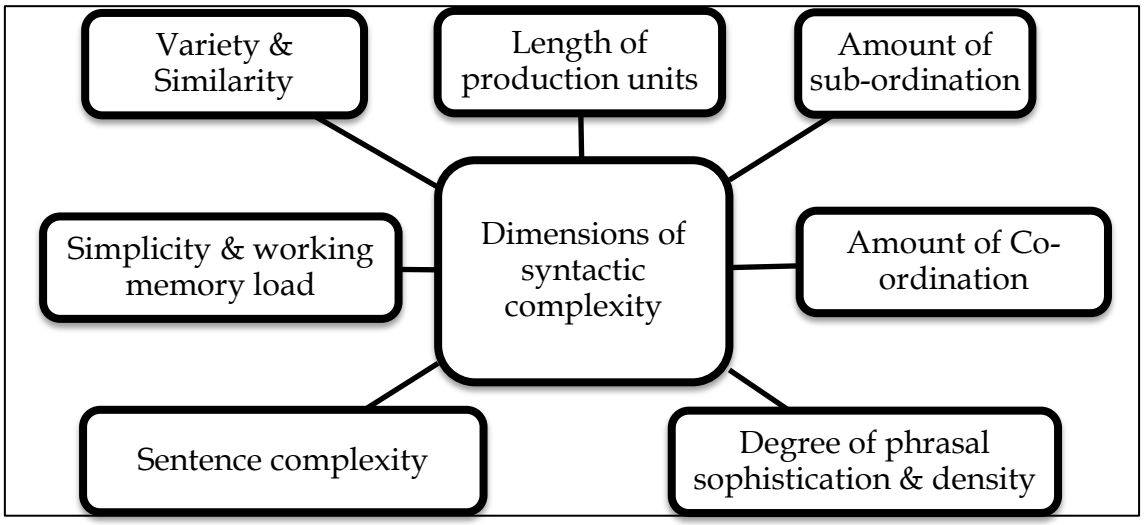
Finally, as mentioned earlier, the study investigated only the *differences* between CEFR levels regarding syntactic complexity by using only *numerical* information (although Section 9.3.2 touched on how the levels could be characterised concerning numerical values obtained for specific SC indices). A more qualitative investigation and characterisation of the CEFR levels concerning SC was not done in this study (see, however, the work done in the English Profile project), and it would be interesting to examine, for example, how the syntax of different L1 background EFL learners might differ qualitatively when their overall writing level on the CEFR scale is equivalent.

9.6 Conclusion

This dissertation contributes to the CEFR-related research on the linguistic foundation of its levels by exploring syntactic complexity differences across the CEFR levels in EFL writing. The study investigated syntactic complexity in English learners' writings from two settings with distinct L1s (Sindhi and Finnish). The research evaluated whether syntactic complexity indices distinguished the CEFR levels A1, A2, B1 and B2 in the two settings. Syntactic complexity in writing has been identified as an essential aspect of language proficiency. It is a multidimensional construct generated from its parts, including length of production units, sentence complexity, coordination, subordination, syntactic similarity, and specific structures (e.g., phrasal structures). The results of the dissertation established considerable differences between the CEFR levels A1, A2, B1, and B2. Additionally, the findings indicated differences in SC in EFL writing between the learners from the two nations (Finland and Pakistan) whose L1 differed. This study examined many elements of syntactic complexity, laying

the groundwork for future research and further work on language teaching, assessment, and curriculum design in Finland and Pakistan.

Figure 9.2 Syntactic complexity dimensions explored in the current dissertation



9.7 Summary

The study explored the development of writing in English as a foreign language, focusing on the relationship between communicative proficiency and mastery of linguistic resources (structures). More specifically, it focused on the differences and similarities in syntactic complexity (SC) across the Common European Framework of Reference (CEFR) levels. The study paves the way for further work on syntactic complexity, for example, analyses of which syntactic features are typical at different CEFR levels.

Syntactic complexity as a multidimensional construct proposed, for example, by Norris and Ortega (2009), was explored in the current dissertation with samples of EFL writing from two diverse linguistic and cultural groups (Sindhis and Finns). Viewing syntactic complexity as a multidimensional construct is a reasonably recent view for examining the SC. It has never been operationalised in CEFR-related research as comprehensively as in the current study, which covered almost 30 indices of SC.

This study compared syntactic complexity between CEFR levels A1, A2, B1 and B2. It further investigated if the difference in SC were related to the EFL learners' native language (L1) and age/educational level. The participants were 8th to 12th-grade EFL learners in Pakistan and 8th and 12th-grade EFL learners in Finland. The study had three sub-studies (see Section 8 for details).

The first research study focused on Pakistani and Finnish students. The study's primary purpose was to compare syntactic complexity across A1, A2, and B1 CEFR levels based on one argumentative text written by the learners. Second, the study sought to determine whether SC varied by L1 (Finnish vs. Sindhi) among EFL learners with the same CEFR writing level. The research questions in Study 1 were: What syntactic elements differentiate CEFR levels A1, A2, and B1 in Sindhi and Finnish EFL argumentative essays? Furthermore, which aspects of syntactic complexity between Sindhi and Finnish EFL learners differ or remain the same when their overall CEFR writing level is the same? Using 30 indices, syntax complexity was examined using two automated analysis tools, L2 Syntactic Complexity Analyzer and Coh-Metrix. In Finland, eighth- and twelfth-grade students' data were analysed, whereas, in Pakistan, eighth- to twelfth-grade students' data were used.

In the second study, 208 eighth graders (aged 14) and 219 upper-secondary school students (aged 17 / 18) from Finland completed three English writing tasks from an international English language exam and a previous study (see Tasks in Section 6.2.2 for details). The tasks required students to narrate, argue, and explain. The first research question was: To what extent does the syntactic complexity in the writing of two age groups of Finnish EFL learners relate to their EFL writing ability? Which SC indices correlate strongest with their ability, and do the two age groups differ? The second research question was: Which SC indices distinguish Finnish EFL learners at different CEFR levels, and do the two age groups differ?

The third study compared two age groups from two countries with the same CEFR proficiency level. Pakistani and Finnish data sets were utilised for this analysis. Whereas Study 1 investigated if EFL learners from different L1 backgrounds differed in terms of syntactic complexity, Study 3 examined the differences in SC between two age (and grade) groups: about 14 years old 8th graders and about 17 years old 12th graders, when the learners' overall CEFR writing level was the same. Although Study 2 included two age groups, it did not directly investigate age-related differences. The research questions in Study 3 were: To what extent does the syntactic complexity of Pakistani EFL learners of different ages with the same CEFR competency level (A1, A2, or B1) differ? To what extent does the syntactic complexity of Finnish EFL learners of different ages with the same CEFR competency level (A1, A2, or B1) differ?

In Pakistan and Finland, the writing tasks were performed under a researcher's or teacher's supervision, adhering to instructions and time limits. Sufficient time was given to the students to write without allowing exhaustion to interfere.

In Pakistan, about 1400 Sindhi-speaking EFL learners from grades 8 to 12 wrote six assignments ranging from e-mail messages to argumentative and narrative texts. The tasks originated from two studies on developing writing skills in Finland: CEFLING (2007-2009) and TOPLING (2010-2013; see sections 6.1.1 and 6.1.2 for details). In Pakistan, the tasks were utilised unaltered or with very slight adjustments. However, only one of the tasks was employed for this

dissertation, namely the one completed by both Finnish student groups (i.e., 8th and 12th graders) in Study 2.

Over 280 8th and 12th-grade EFL learners in Finland authored descriptive, argumentative, and narrative texts (approximately 1280 scripts). The eighth- and twelfth graders were assigned two separate tasks and one shared task; the Pakistani students also completed this last task. The first and second tasks assigned to the Finnish groups were from the PTE General as Pearson and the Dialuki project collaborated (see Alderson et al. 2015 and Alderson, Huhta & Nieminen 2016 for details).

The rating data were analysed with multifaceted Rasch analysis (with the programme Facets) to ensure that the placement of the learner scripts on the CEFR levels was consistent enough. Then, the data (SC indices) from the L2 Syntactic Complexity Analyzer and Coh-Metrix were analysed with SPSS. MANOVA, correlations, T-tests, and descriptive statistics determined if syntactic complexity was related to CEFR levels A1, A2, B1, and B2 and if the levels differed regarding SC among the study's learner groups. In each syntactic complexity index group, Mahalanobis distance tests were used to examine and eliminate multivariate outliers for each dimension of SC. To analyse CEFR level differences, multivariate analyses of variance were then performed for each dimension or combination of indices (e.g., L2SCA count variables or L2SCA SC variables. Then, univariate analysis and pairwise comparisons were utilised to determine indices that distinguished CEFR levels. The link between SC indices and writing skills was analysed using Pearson correlation coefficients (learner ability measures from Facets were used as indicators of writing skill).

In general, the findings of this research help comprehend the relationship between SC characteristics and skill levels, as well as the development of SC as learners advance. CEFR levels were discovered using multivariate research. The length of the production unit and subordination indices differed between CEFR levels in L1 and age cohorts. Several differences between the CEFR levels and between the Pakistani and Finnish EFL learners were found, particularly in the length of production units and phrasal level indices, working memory load, referring expressions, and syntactic diversity and simplicity.

The study has, thus, demonstrated that syntactic complexity in EFL learners' writing generally rises as the learners' writing ability increases (at least in the CEFR level A1 to B1/B2 range) and that many indices of SC significantly differentiate between the CEFR levels, as has been found in several previous studies. The has also shown that EFL learners' first language background matters in that, for example, Sindhi-speaking and Finnish-speaking EFL learners' syntax differs even if their overall CEFR writing level is the same. Furthermore, the EFL learners' age (or grade level) affects their syntactic complexity when the learners' CEFR level is the same, and different L1 groups can differ in this respect. Such L1 or age-related matters have yet to be studied before.

The current quantitative study investigated syntactic complexity indices only in terms of their numerical values, focusing on differences between CEFR level and L1 and age groups. Further research is thus needed to shed light on

which specific syntactic features are typical of CEFR levels, allowing them to be characterised in more qualitative terms.

REFERENCES

Agard, Frederick B. Dunkel Harold B. (1948). *An investigation of Second Language Teaching*. Published by Ginn and company, the university of Chicago.

Ahola, S., & Leblay, T. (2014). Kaksi vuosikymmentä Yleisiä kielitutkintoja. *Kieli, koulutus ja yhteiskunta*, *18*(Joulukuu).

Ai, H., & Lu, X. (2013). *A corpus-based comparison of syntactic complexity in NNS and NS university students' writing* (pp. 249–264). https://doi.org/10.1075/scl.59.15ai

Alanen, R., Huhta, A., & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. *In I. Bartning, M. Martin & I. Vedder (Eds.), Communicative proficiency and linguistic development: Intersections between SLA and language testing research (pp. 21–56). Eurosla Monographs Series, 1. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html*.

Alderman, L. R., (1927). *Public Education of adults in the years 1924-1926*. United States Government Printing Office Washington.

Alderson, J. C. (2007). The CEFR and the need for more research. In *Modern Language Journal* (Vol. 91, Issue 4). https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2014). *The diagnosis of reading in a second or foreign language*. Routledge.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, *22*(3). https://doi.org/10.1191/0265532205lt310oa

Alderson, J. C., Huhta, A., & Nieminen, L. (2016). Characteristics of weak and strong readers in a foreign language. *The Modern Language Journal*, *100*(4), 853-879.

Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2014). *The diagnosis of reading in a second or foreign language*. Routledge.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning*, *67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Anwar, B. (2009). Urdu-English code switching: The use of Urdu phrases and clauses in Pakistani English (A non-native variety). *International Journal of Language Studies*, *3*(4).

Arnaud, P. J. L. 1992. 'Objective lexical and component tests' in P. J. L. Arnaud and grammatical characteristics of 12 written H. Bejoint (eds): Vocabulary and Applied compositions and the validity of separate-component tests' in P. J. L. Arnaud and H. Bejoint (eds): *Vocabulary and Applied Linguistics. London*: Macmillan. pp. 133-45.

Arnold, T., Ballier, N., Gaillat, T., & Lissón, P. (2018). Predicting CEFRL levels in learner English on the basis of metrics and full texts. *Proceedings of the 20th*

Conférence Sur l'Apprentissage Automatique. INSA de Rouen, 20–22 June. https://arxiv.org/pdf/1806.11099.pdf

Arthur, B. 1979. 'Short-term changes in EFL composition skills" in C. Yorio, K. Perkins, and J. Schachter (eds): *On TESOL '79: The Learner in Focus. Washington*, DC: *TESOL. pp*. 330-342.

Asghar, Z. M., Mahmood, M. A., & Asghar, S. A. (2014). A Corpus Based Study of Distribution of Preposition in Pakistani and British Englishes. *Journal of Education and Practice*, 213-224.

Bachman, L. F., & Cohen, A. D. (1998). Language testing–SLA interfaces: An update. *Interfaces between second language acquisition and language testing research*, 1-31.

Banerjee, J., F. Franceschina, and A. Smith (2007). 'Documenting features of written language production typical of different IELTS band score levels,' IELTS Research Reports, vol. 7. IELTS Australia and British Council.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11, 17–34. doi:10.1017/S0272263100007816.

Barrot, J. S., & Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, *47*, 100510. https://doi.org/10.1016/j.asw.2020.100510

Bartning, I., Martin, M., & Vedder, I. (2010). *Communicative proficiency and linguistic development* (Vol. 1). Lulu. com.

Baumgardner, R. J. (1990). The indigenization of English in Pakistan. *English today*, *6*(1), 59-65.

Baumgardner, Robert J. (1993a). The Engish language in Pakistan, 41-54. Karachi: Oxford University Press.

Baumgardner, Robert J. (1993b). The indigenization of English in Pakistan. In: Robert J. Baumgradner (ed.), The Engish language in Pakistan, 41-54. Karachi: Oxford University Press.

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing*, *22*(2). https://doi.org/10.1007/s11145-007-9107-5

Bernstein Ratner, N. (2004). Fluency and stuttering in bilingual children. *Bilingual language development and disorders in Spanish-English speakers*, 287-308.

Bhat, S., & Yoon, S. Y. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, *67*, 42-57.

Biber, D. (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. , *23*(2), 337 360. https://doi.org/10.1515/ling.1985.23.2.337

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, Vol. 62, No. 2 (Jun., 1986), pp. 384-414.

Biber, D (1988). VARIATION ACROSS SPEECH AND WRITING. Cambridge: Cambridge University Press, 1988. Pp. xiii+ 299. $49.50.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, *15*(2), 133–163. https://doi.org/10.1080/01638539209544806

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, *46*. https://doi.org/10.1016/j.jeap.2020.100869

Bloomfield Leonard (1926). *Language.* Published by George Allen and Uniwin Ltd London.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). https://doi.org/10.1075/lllt.32.02bul

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics (United Kingdom)*, *28*(1). https://doi.org/10.1111/ijal.12196

Bulté, B., & Roothooft, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, *91*. https://doi.org/10.1016/j.system.2020.102246

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

Carlsen, C. (2012). Proficiency Level--a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics*, *33*(2), 161–183. https://doi.org/10.1093/applin/amr047

Casanave, C. 1994. 'Language development in students' journals." *Journal of Second Language Writing* 3: 179-201.

Caplan, D., Michaud, J., & Hufford, R. (2013). Dissociations and associations of performance in syntactic comprehension in aphasia and their implications for the nature of aphasic deficits. *Brain and language*, *127*(1), 21-33.

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 722-731).

Chen, Y. H., & Baker, P. (2014). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. In *Applied Linguistics* (Vol. 37, Issue 6). https://doi.org/10.1093/applin/amu065

Condouris K, Meyer E, Tager-Flusberg H (2003). The relationship between standardized measures of language and measures of spontaneous speech

in children with autism. *Am J Speech Lang Pathol*. 2003 Aug;12(3):349-58. doi: 10.1044/1058-0360(2003/080). PMID: 12971823; PMCID: PMC1201513.

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research*, *69*(5), 176-183.

Council of Europe (CoE), (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. CUP https://rm.coe.int/16802fc1bf

Council of Europe (CoE), (2020). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume. Council of Europe. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, *42*(3), 475-493.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The Development of Writing Proficiency as a Function of Grade Level: A Linguistic Analysis. *Written Communication*, *28*(3), 282–311. https://doi.org/10.1177/0741088311410188

Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015, March). Pssst... textual features... there is more to automatic essay scoring than just you!. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 203-207).

DeKeyser, R.M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty, & J. Williams (Eds.). Focus on form in classroom language acquisition (pp. 42–63). Cambridge: Cambridge University Press.

DeKeyser, R.M. (2005). What makes learning second-language grammar difficult? A review of issues. Language Learning, 55 (Suppl. 1), 1–25.

Ellis, R. (1990). Instructed second language acquisition. Oxford: Blackwell.

Ellis, R. (1994). The study of second language acquisition. Oxford: Oxford University Press.

Ellis, R. (2003). Task-based language learning and teaching. Oxford: Oxford University Press.

Ellis, R. (2008). The study of second language acquisition (2nd ed.). Oxford: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005a). Analysing learner language. Oxford: OUP.

Ellis, N., & Robinson, P. (2008). An introduction to cognitive linguistics, second language acquisition, and language instruction. In P. Robinson, & N. Ellis (Eds.). Handbook of cognitive linguistics and second language acquisition (pp. 3–24). New York: Routledge.

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.

Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.). Planning and task performance in a second language (pp. 167–192). Amsterdam: John Benjamins.

Engelhard, G. (1994). Examining Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Erickson, G., Åberg-Bengtsson, L., & Gustafsson, J. E. (2015). Dimensions of test performance in English as a foreign language in different European settings: a two-level confirmatory factor analytical approach. *Educational Research and Evaluation*, *21*(3), 188-208.

Ferris, M. R., & Politzer, R. L. (1981). Effects of early and delayed second language acquisition English composition skills of Spanish-speaking junior high school students. *TESOL Quarterly*, *15*(3), 263-274.

Flahive, D. E. and B. G. Snow. 1980. 'Measures of syntactic complexity in evaluating ESL compositions" in J. W. Oller and K. Perkins (eds): *Research in Language Testing*. Rowley, MA: Newbury House. pp. 171-176.

Frederick Browning Agard, Harold Baker Dunkel (1948). *An Investigation of Second-language Teaching, Investigation of the Teaching of a Second Language*. University of Chicago, Ginn.

Gaies, S. J. 1976. Sentence-combining: A technique for assessing proficiency in a second language. *Paper delivered at the Conference on Perspectives on Language, University of Louisville*, Kentucky.

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouvé, M., & Zarrouk, M. 2021. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 1-17. https://doi.org/10.1017/S095834402100029X

Galaczi, E. D., & Weir, C. J. (Eds.). (2013). *Exploring language frameworks: proceedings of the ALTE Kraków Conference, July 2011* (Vol. 36). Cambridge University Press.

Garam, I. (2009). Degree programmes taught through a foreign language in Finnish higher education. *Faktaa: Facts and Figures, 2B*.

Gardezi, S. A., & Nesi, H. (2009). Variation in the writing of economics students in Britain and Pakistan: The case of conjunctive ties. *Academic writing: At the interface of corpus and discourse*, 236-250.

Gipps, C., & Ewen, E. (1974). Scoring Written Work in English as a Second Language the Use of the T-unit. *Educational Research*, *16*(2), 121-125.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004a). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research*

Methods, Instruments, & Computers, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Grainger, J. M. (1907). *Studies in the syntax of the King James version* (Vol. 2). University Press.

Green, A. 2012. *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition across the Ability Range*. CUP.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR. *EUROSLA Yearbook*, *14*, 1–30. https://doi.org/10.1075/eurosla.14.01gyl

Halleck, G.B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *Modern Language Journal*, 79(2), 223–234.

Harding, L., Alderson, J.C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, 32(3), 317–336.

Härmälä, M., Huhtanen, M., Puukko, M., & Marjanen, J. (2019). A-englannin oppimistulokset 7. luokan alussa 2018. *Julkaisut*, *13*, 2019.

Härmälä, M., Huhtanen, M., Puukko, M., & Marjanen, J. (2014). Englannin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013. https://karvi.fi/publication/englannin-kielen-oppimaaran-oppimistulokset-perusopetuksen-paattovaiheessa-2013/

Hawkins, J., and L. Filipovic´. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. CUP.

Henry, K. 1996. 'Early L2 writing development: A study of autobiographical essays by university-level students of Russian." *The Modern Language Journal* 80: 309-26.

Hirano, K. 1991. 'The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students.' *Annual Review of English Language Education in Japan* 2: 21-30.

Homburg, T. J. 1984. 'Holistic evaluation of ESL compositions: Can it be validated objectively?" *TESOL Quarterly* 18: 87-107.

Ho-Peng, L. 1983. 'Using T-unit measures to assess writing proficiency of university ESL students." *RELC Journal* 14: 35-43.

Hoppes, W. C. (1934). Considerations in the Development of Children's Language. *The Elementary English Review*, *11*(3), 66-70.

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

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

Huhta, A. (2012). Common European Framework of Reference. In *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0157

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing*, *31*(3). https://doi.org/10.1177/0265532214526176

Hulstijn, J. H. (2007a). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. In *Modern Language Journal* (Vol. 91, Issue 4). https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

Hulstijn, J. H. (2007b). Fundamental issues in the study of second language acquisition. *EUROSLA Yearbook*, *7*, 191–203. https://doi.org/10.1075/eurosla.7.11hul

Hulstijn, J., Alderson, J. C., & Schoonen, R. (2010). 'Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them'. In Bartning, I., Martin, M., Vedder, I. Eds., *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research* pp. 11–20. Eurosla Monographs Series 1.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of dutch as a second language at the B1 and B2 levels of speaking proficiency of the common european framework of reference for languages (CEFR) 1. *Language Testing*, *29*(2). https://doi.org/10.1177/0265532211419826

Hunt, K.W. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3. Champaign, IL: National Council of Teachers of English.

Hunt, K. W. (1966). Sentence Structures Used by Superior Students in Grades Four and Twelve, and by Superior Adults.

Hunt, K. W. (1968). An instrument to measure syntactic maturity, preliminary version. Florida st. Univ., tallahassee report number br -6 -225

Hunt, K. W. (1969). Another Probe into Syntactic Maturity. Louisiana English Journal, 9(1), 5-15.

Hunt, K.W. (1970). Syntactic maturity in school children and adults. Monograph of the Society for Research in Child Development, 134(35), 1.

Hunt, K. W., & O'donnell, R. (1970). An Elementary School Curriculum to Develop Better Writing Skills.

Hynninen, N. (2016). *Language regulation in English as a lingua franca: Focus on academic spoken discourse* (Vol. 9). Walter de Gruyter GmbH & Co KG.

Ishikawa, S. 1995. 'Objective measurement of low-proficiency EFL narrative writing." *Journal of Second Language Writing* 4: 51-69.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 1–23.

Jagaiah, T. (2017). Analysis of Syntactic Complexity and its Relationship to Writing Quality in Argumentative Essays (Unpublished doctoral dissertation). *University of Connecticut, Mansfield, CT, United States*.

Jagaiah, T., Olinghouse, N. G., & Kearns, D. M. (2020). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. In *Reading and Writing* (Vol. 33, Issue 10). https://doi.org/10.1007/s11145-020-10057-x

Jespersen, Otto (1905). *Growth and Structure of the English language.* Published by B. G. Teubner Leipzig, 474 Pages.

Jenkins, J., Baker, W., & Dewey, M. (Eds.). (2018). *The Routledge handbook of English as a lingua franca* (pp. xix-620). London: Routledge.

Jiang, J., Bi, P., & Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, *46*. https://doi.org/10.1016/j.jslw.2019.100666

Jo, C. W. (2021). Short vs. extended adolescent academic writing: A cross-genre analysis of writing skills in written definitions and persuasive essays. *Journal of English for Academic Purposes*, *53*. https://doi.org/10.1016/j.jeap.2021.101014

Jo, C. W. (2022). Mapping adolescent literacy across L1 backgrounds: Linguistic and discourse features as predictors of persuasive essay quality. *System*, *104*. https://doi.org/10.1016/j.system.2021.102698

Kameen, P. 1979. 'Syntactic skill and ESL writing quality" in C. Yorio, K. Perkins, and J. Schachter (eds): *On TESOL* 79: The Learner in Focus. Washington, DC: *TESOL*. pp. 343—64.

Kaplan, R. B., & Baldauf Jr, R. B. (2005). Language planning and policy in Europe: Vol I: Hungary, Finland and Sweden.

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across common european framework of reference levels A1, A2, and B1. *Applied Linguistics*, *41*(4). https://doi.org/10.1093/applin/amy064

Khushik, G. A., & Huhta, A. (2022). Syntactic complexity in Finnish-background EFL learners' writing at CEFR levels A1–B2. *European Journal of Applied Linguistics*, *10*(1), 142–184. https://doi.org/10.1515/eujal-2021-0011

Kim, M. (2021). Exploring longitudinal changes in lexical and syntactic features in beginning-level EFL learner writing. *System*, *103*. https://doi.org/10.1016/j.system.2021.102680

Kim, S. 2004. 'A study of development in syntactic complexity by Chinese learners of English and its implications on the CEF scales', MA dissertation, Lancaster University.

Korhonen, K. (2010). Teaching English in Finnish upper secondary schools: How English is taught and what are the most effective ways of learning.

Kortmann, B., & Szmrecsanyi, B. (2004). Global synopsis: morphological and syntactic variation in English. *A handbook of varieties of English*, *2*, 1142-1202.

Kroll, B. (1990). *Second language writing* (Vol. 10). Cambridge: Cambridge University Press.

Krook, A. (1893). The english language in Finland. Helsingfors: Société néophilogique de Helsinki.

Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. *Eurosla yearbook*, *5*(1), 195-222.

Kuiken, F., Vedder, I., & Matters, M. (2007). Cognitive task complexity and linguistic performance in French L2 writing. *Investigating tasks in formal language learning*, 117-135.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. http://scholarworks.gsu.edu/alesl_diss

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513-535.

Kyle, K., & Crossley, S. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *Modern Language Journal*, *102*(2). https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, *43*(4). https://doi.org/10.1017/S0272263120000546

Lahuerta Martínez, A. C. (2018). Analysis of syntactic complexity in secondary education ELF writers at different proficiency levels. *Assessing Writing*, *35*. https://doi.org/10.1016/j.asw.2017.11.002

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

Larsen-Freeman, D. 1976. 'Evidence of the need for a second language acquisition index of development" in W. C. Ritchie (ed.): *Principles of Second Language Learning and Teaching*. New York: Academic Press.

Larsen-Freeman, D. 1983. 'Assessing global second language proficiency' in H. W. Seliger and M. H. Long (eds): *Classroom Oriented Research in Second Language Acquisition*. Rowley, MA: Newbury House. pp. 287—304. [Study 2, cross-sectional data; and Study 3, longitudinal data.]

Larsen-Freeman, D. and V. Strom. 1977. "The construction of a second language acquisition index of development." *Language Learning* 27: 123-34.

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.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 579–589.

Larsen-Freeman, D., & Cameron, L. (2008). Complex systems and *applied linguistics*. Oxford: Oxford University Press.

Larsson, T., & Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes*, *45*. https://doi.org/10.1016/j.jeap.2020.100850

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, 32(3), 299–316.

Lee, C., Ge, H., & Chung, E. (2021). What linguistic features distinguish and predict L2 writing quality? A study of examination scripts written by adolescent Chinese learners of English in Hong Kong. *System*, *97*. https://doi.org/10.1016/j.system.2021.102461

Linacre, J. M. (1989). Many faceted Rasch measurement. Unpublished doctoral dissertation, University of Chicago.

Linacre, J. M. (2009). Facets Rasch measurement computer program. Chicago: Winsteps.com.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing*, *22*(3). https://doi.org/10.1191/0265532205lt311oa

Littlemore, J., Krennmayr, T., Turner, J., & Turner, S. (2014). An investigation into metaphor use at different levels of second language writing. *Applied linguistics*, *35*(2), 117-144.

Low, W. H and Lond, M. A (1897). *English Language: Its history and structure.* London: W. B Clive, University of Correspondence college press. Warehouse: 13 Booksellers Row, Strand, W.C.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4). https://doi.org/10.1177/0265532217710675

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*. https://doi.org/10.1016/j.jslw.2015.06.003

Mahboob, A. (2003). The English language in Pakistan: A brief overview of its history and linguistics. *Pakistan Journal of Language*, *4*(1), 1-28

Mahboob, A. (2015a) Identity Management, Language Variation, and English Language Textbooks. In Djenar, D., Mahboob, A. & Cruickshank, K. (Eds.) Language and Identity Across Modes of Communication. Boston: Mouton de Gruyter.

Mahboob, A. (2015b). Understanding and providing 'cohesive' and 'coherent' feedback on writing. Writing and Pedagogy, Vol 7.2

Mahmood, M. A., & Ali, N. (2011). Circumstance adverbials in Pakistani English: A corpus-based comparative analysis. *International journal of academic research*, *3*(5), 47-51.

Mahmood, R. (2009). Lexico-Grammatical study of noun phrase in Pakistani English. *Unpublished doctoral dissertation). Bahauddin Zakariya University, Multan*.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of second language writing*, *29*, 3-15.

Matsuda, P. K., & De Pew, K. E. (2002). Early second language writing: An introduction. *Journal of Second Language Writing*, *4*(11), 261-268.

McCarthy, P. M., Guess, R. H., & McNamara, D. S. (2009). The components of paraphrase evaluations. *Behavior Research Methods*, *41*(3), 682–690. https://doi.org/10.3758/BRM.41.3.682

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching*, *44*(4). https://doi.org/10.1017/S0261444811000073

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. https://doi.org/10.1177/0265532211430367

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292-330.

Miestamo, M., Sinnemäki, K., & Karlsson, F. (Eds.). (2008). *Language complexity: Typology, contact, change*. Amsterdam: John Benjamins.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2.

Miles, S., & Bernstein Ratner, N. (2001). Parental language input to children at a stuttering onset. *Journal of Speech, Language & Hearing Research*, 44, 1116–1130

Miller, D. W., & Toman, M. (2015). An analysis of the syntactic complexity in service corporation brand slogans. *Services Marketing Quarterly*, *36*(1), 37-50.

MOEC, (2018). Ammattikorkeakoulu-uudistuksen arviointi. Loppuraportti accessed from https://julkaisut.valtioneuvosto.fi/handle/10024/161050

Monroe, J. H. 1975. 'Measuring and enhan-cing syntactic fluency in French." The French Review 48: 1023-31.

Mukherjee, J., & Huber, M. (2015). *Corpus linguistics and variation in English: Theory and description*. BRILL.

Nayyar, A. H., & Salim, A. (2005). The subtle subversion: The state of curricula and textbooks in Pakistan Urdu, English, Social Studies and Civics. Sustainable Development Policy Institute.

Neff-van Aertselaer, J. (2013). Contextualizing EFL argumentation writing practices within the Common European Framework descriptors. *Journal of Second Language Writing*, *22*(2), 198-209.

Neff, J., E. Dafouz, M. Diez, R. Prieto, and C. Chaudron. 1998. Contrastive discourse analysis: Argumentative text in English and Spanish. Paper presented at the Twenty- Fourth Linguistics Symposium: Discourse across Languages and Cultures, University of Wisconsin-Milwaukee, September.

Nichols, J. (2009). Linguistic complexity: A comprehensive definition and survey. In *Language complexity as an evolving variable* (pp. 110-125). Oxford University Press.

Nihalani, N. K. 1981. 'The quest for the L2 index of development." *RELC Journal* 12: 50—6.

Nikula, T., Pöyhönen, S., Huhta, A., & Hildén, R. (2010). When MT+ 2 is not enough: Tensions within foreign language education in Finland. *Sociolinguistica: Internationales Jahrbuch für Europäische Soziolinguistik= International Yearbook of European Sociolinguistics= Annuaire International de la Sociolinguistique Européenne*, (24), 25-42.

Norris, J. M. and L. Ortega. 2000. 'Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis." *Language Learning* 50: 417-528.

Norris, J. and Ortega, L. **(**2009). 'Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA'. *Applied Linguistics, 30,* 555–578.

North, Brian. 2020. *The CEFR illustrative descriptors: Validation reference paper for researchers* [Unpublished manuscript]. Council of Europe. https://www.researchgate.net/publication/350156412_North_CEFR_Validation_Reference_Paper_-_with_appendices (accessed 15 July 2021).

North, B. (2021). The CEFR Companion Volume—What's new and what might it imply for teaching/learning and for assessment?. *CEFR Journal–Research and practice*, *4*, 5-24.

O'Donnell, R. C. (1967). Syntax of Kindergarten and Elementary School Children: A Transformational Analysis. *National Council of Teachers of English*, 1111 Kenyon Road, Urbana, Ill.

O'DONNELL, R. C. (1968). An objective measure of structural complexity in children's writing. *National Council of Teachers of Engl*ish, 1111 Kenyon Road, Urbana, Ill.

OECD (2015). Indicators Education at a Glance 2015: Finland more info: https://doi.org/10.1787/eag-2015-55-en

OECD **(2**016). *PISA 2015 Results (Volume I): Excellence and Equity in Education.* PISA, OECD Publishing.

OECD (2018). Programmae for international student assessment results from PISA (2018). Accessed from https://www.oecd.org/pisa/publications/PISA2018_CN_FIN.pdf

O'Shea, M. V. (1907). *Linguistic development and education*. Macmillan.

OPH (2019). Impact program work 2019. Accessed from website https://www.oph.fi/en/node/226

OPH (2003 and 2004). Perusopetuksen opetussuunnitelman perusteet. accessed from https://www.oph.fi/fi/koulutus-ja-tutkinnot/perusopetuksen-opetussuunnitelman-perusteet.

OPH (2014 and 2015). Perusopetuksen opetussuunnitelman perusteet. accessed from https://www.oph.fi/fi/koulutus-ja-tutkinnot/perusopetuksen-opetussuunnitelman-perusteet

OPH (2020). Perusopetuksen opetussuunnitelman perusteet. accessed from https://www.oph.fi/fi/koulutus-ja-tutkinnot/perusopetuksen-opetussuunnitelman-perusteet

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*. https://doi.org/10.1016/j.jslw.2015.06.008

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates Publishers. https://psycnet.apa.org/record/2003-02475-003

Palmer Harold E. (1921). The principles of langauge study. Yonkers on Hudson, New York, World Book Company, 184 pages.

Palviainen, Å., Kalaja, P., & Mäntylä, K. (2012). Development of L2 writing: Fluency and proficiency. *AFinLA-e: Soveltavan kielitieteen tutkimuksia*, (4), 47-59.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second language research*, *35*(1), 121-145.

Peelle, J. E., McMillan, C., Moore, P., Grossman, M., & Wingfield, A. (2004). Dissociable patterns of brain activity during comprehension of rapid and syntactically complex speech: evidence from fMRI. *Brain and language*, *91*(3), 315-325.

Perkins, K. 1980. 'Using objective methods of attained writing proficiency to discriminate among holistic evaluations." *TESOL Quarterly*, 14: 61-7.

Perkins, K. and T. J. Homburg. 1980. "Three different statistical analyses of objective measures of attained writing proficiency" in R. Silverstein (ed.): *Occasional Papers in Linguistics*, No. 6. Carbondale, IL: Southern Illinois University. pp. 326-37.

Polat, N., Mahalingappa, L., & Mancilla, R. L. (2020). Longitudinal growth trajectories of written syntactic complexity: The case of turkish learners in an intensive english program. *Applied Linguistics*, *41*(5). https://doi.org/10.1093/applin/amz034

Polio, C. 0. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning,* 47, 101—14

QI, D. (2014). Syntactic complexity of EFL, ESL and ENL: Evidence from the international corpus network of Asian learners of English.

Rahman, T**.** 2001. 'English-teaching institutions in Pakistan'. *Journal of Multilingual and Multicultural Development,* 22/3, 242-261.

Rahman, T. (2016). The development of English in Pakistan. In G. Leitner, A. Hashim, & H. Wolf (Eds.), Communicating with Asia: The Future of English as a Global Language (pp. 13-27). Cambridge: Cambridge University Press. doi:10.1017/CBO9781107477186.002

Ratner, N. B. (1988). Patterns of parental vocabulary selection in speech to very young children. *Journal of Child Language*, *15*(3), 481-492.

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

Ringbom, H. (2012). Review of recent applied linguistics research in Finland and Sweden, with specific reference to foreign language learning and teaching. *Language Teaching*, *45*(4), 490-514.

Robinson, P. (2002). Individual differences and instructed language learning. *Individual Differences and Instructed Language Learning*, 1-399.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design.

Robinson, P., & Ellis, N. C. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition* (Vol. 270). New York: Routledge.

Rosmawati, M. (2019). Syntactic Complexity in Second Language Academic Writing in English: Diversity on Display. *Voices and practices in applied linguistics: Diversifying a discipline*, 251-270.

Ryshina-Pankova, M. (2015). A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor. *Journal of Second Language Writing*, *29*, 51-63.

Saari, H., & Takala, S. (1980). English as a Foreign Language in Finland. *Studies in Educational Evaluation*, *6*(1), 47-52.

Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian modern language review*, *68*(2), 138-163.

Shamim, F. 2008. 'Trends, issues and challenges in English language education in Pakistan', *Asia Pacific Journal of Education*, 28/3, 235-249.

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

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

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task based learning. Language Teaching Research, 1(3), 185–211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. Language Learning, 49(1), 93–120.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.). Cognition and second language instruction (pp. 183–205). Cambridge: Cambridge University Press.

Smith, C. A. (1893). The order of words in Anglo-Saxon prose. *PMLA*, *8*(2), 210-244.

Smith, C. A. (1904). The Short Circuit in English Syntax. *Modern Language Notes*, *19*(5), 113-121.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*, 31, 532–553.

Storch, N., & Wigglesworth, G. (2007). Writing tasks: The effects of collaboration. In M.P. García Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 157–177). Clevedon: Multilingual Matters.

Talaat, M. (2002). The Sociolinguistics of English in Pakistan: Form and Functions. Unpublished Dissertation, Bahauddin Zakria University, Multan.

Talaat, M. (2003). Pakistani English: A sociolinguistic variety. *Journal of Research (Faculty of Languages & Islamic Studies)*, *4*.

Talbot Winthrop (1916). *Adult literacy*. Bulletin 1916, No. 35. Washington Government Printing Office.

Tomita, Y. (1990). T-unit omochiita kokosei no jiyu eisaku bun noryoku no sokutei (Assessing the writing ability of high school students with the use of T-units). *Step Bulletin*, 2, 14—28

Torrance, M. , A. D. , C. M. , K. O. , M. A. , T. L. , & V. W. L. (2012). *Learning to Write Effectively: Current Trends in European Research* (M. Torrance, D. Alamargot, M. Castelló, F. Ganier, O. Kruse, A. Mangen, L. Tolchinsky, & L. van Waes, Eds.). BRILL. https://doi.org/10.1163/9781780529295

Trumpp, E. (1872). *Grammar of the Sindhi Language Compared with the Sanskrit-Prakrit and the Cognate Indian Vernaculars Ernes Trumpp*. Trubner and C..

Vann, R. J. (1979). Oral and written syntactic relationships in second language learning. *On Tesol*, *79*, 322-329.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239–263. https://doi.org/10.1016/j.jslw.2012.03.007

Wächter, B., & Maiworm, F. (2008a). English-taught programmes in European higher education. *ACA Papers on International Cooperation in Education. Bonn: Lemmens*.

Wächter, B. (2008b). Teaching in English on the rise in European higher education. *International higher education*, (52).

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3). https://doi.org/10.1191/0265532205lt309oa

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis* (Vol. 43). John Wiley & Sons.

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. In *Modern Language Journal* (Vol. 91, Issue 4). https://doi.org/10.1111/j.1540-4781.2007.00627_9.x

Williams, J., & Evans, J. (1998). What kind of focus and on which forms? In C. Doughty, & J. Williams (Eds). *Focus on form in classroom second language acquisition* (pp. 139–151). Cambridge: Cambridge University Press.

Wisniewski, K. (2017). Empirical Learner Language and the Levels of the Common European Framework of Reference. In *Language Learning* (Vol. 67). https://doi.org/10.1111/lang.12223

Wolfgang, C. H. (1995). *Solving discipline problems: Methods and models for today's teachers*. Prentice Hall. https://doi.org/10.1016/j.jeap.2019.100798

Wyld, H. C (1907). *The growth of English. An elementary account of the present form of our language, and its development*. Published by New York E. P., Dutton and Company.

Ylönen, S., & Kivelä, M. (2011). The role of languages at Finnish universities. *Apples-Journal of Applied Language Studies*.

Appendix 1. Table 1

Studies on syntactic complexity till Millennium (non- CEFR studies, i.e., studies not referring to the CEFR levels. Studies are cross-sectional)

The list of abbreviations: W/C = words per clause, W/S = words per sentence, W/T = words per T-unit, T/S = T-unit per sentence, C/T = clauses per T-unit, C/S = clauses per sentence, DC/C = dependent clauses per clause, DC/T-unit = dependent clause per T-unit, CN/T = complex nominals per T-unit.

In the table, *** denotes large effect size, ** moderate effect size, * small effect size, and (X) means non-significant results

Study	Writers / Participants / L1, L2, FL (languages) / # of levels / Proficiency	Samples / Analysis	(***) = Significant results with large effect size / (**) = Results with moderate effect size. / (*) = Results with small effect size. / (X) = Non-significant results
Gipps & Ewen, (1974)	751 school aged children / L1 various (Asian) / SL English / 3 groups defined by years of L2 schooling	first 100-words from essays on one of 3 given topics / Statistical comparison between groups on measures (no specifics provided)	Words per T-unit (**)
Monroe (1975)	110 US / FL / 4 university course levels & NS rewriting	re-writing simplified passage /alludes to an ANOVA effect of levels on measures	W/C, W/S, W/T, T/S, C/T (**).
Cooper (1976)	50 university students /L1 English, FL German / Four course levels and Native writers	500 words from essays on various topics / MANOVA effect of levels on measures	W/T, W/C, W/S, C/T, CN/T (**). CP/T (*). T/S (X).
Gaies (1976)	NS (16) NNS (5) NNS (20) / L2 English VS natives / TOEFL (4 levels)	Re-write Aluminum paragraph / Correlation between measures and TOEFL scores	MLT, MLC, C/T (*).
Larsen-freeman & Storm (1977)	48 University Students /L1 various, L2 English / 5 levels (Holistic judgment of writing)	essays on given topic / ANOVA effect of levels on measures	W/T (X).
Larsen-freeman (1978)	212 US / L1 various, L2 English / 5 program levels	30-minute essays on one topic / ANOVA effect of levels on measures	W/T (**).
Arthur, (1979) study 2	14 university students / L1 various / low intermediate, intact	essays on one given topic, no reported time limit / multiple regression	W/T (X).

	class holistic rank ordering of essays by teachers	discriminant analysis correlating rankings with measures	
Kameen (1979)	50 US / L1 various, L2 English / 2 levels derived from holistic ratings of writing samples for a program placement test	30-minute essays on various topics / Wilcoxon signed rank tests comparing good & poor groups on measures	W/C, W/T, W/S (**). DC/C, C/T (X).
Vann (1979) analysis 1	28 graduate students / L1 Arabic, SL English / TOEFL scores, 3 levels based on holistic ratings	20-minute written summary and response to a film / analysis 1: multiple regression step analysis correlating TOEFL with measures	W/T, DC/T (X).
Vann (1979) analysis 2	28 graduate students / L1 Arabic, SL English / TOEFL scores, 3 levels based on holistic ratings	20-minute written summary and response to a film / comparison of low and high holistic groups on measures	W/T, DC/T (X).
Perkins (1980)	29 US / L1 various, SFL English / Advanced learners on Michigan test (mean: 74.92). Holistic ratings (Pass, Pass-, Fail)	50-minute essays on one of 3 given topics / ANOVA on effect of holistic ratings on measures	W/T, C/T (X).
Sharma (1980)	60 US / L1 various, SFL English / 3 program levels based on Michigan test	Re-writing aluminium / t-tests (intermediate and advanced levels)	W/C, W/T, C/T (X).
Flahive & Snow (1980) analysis 1	300 US / L1 various, SFL English / 6 program levels, holistic ratings writing within each level based on 5-point scale	50-minute essay on one of several given topics / Discriminant analyses of 3 collapsed groups.	W/T (***). C/T (*).
Flahive & Snow (1980) analysis 2	300 US / L1 various, SFL English / 6 program levels, holistic ratings writing within each level based on 5-point scale	50-minute essay on one of several given topics / Second analyses: correlation between holistic rating & measures for each level	W/T (Highest) (***). C/T Level 1 (lowest), C/T (Level 2), C/T (Level 3), W/T (Level 4), W/T (Level 5), C/T Level 6 (Highest) (**). W/T (Level 2), W/T (Level 3), C/T (Level 5) (*). W/T Level 1 (lowest), C/T (Level 4) (X).
Nihalani (1981)	29 US / L1 Indian languages, FL English / Holistic judgments of writing samples	Take home essays / ANOVA on effect of level on measures	W/T (X).

Ho-Peng (1983) Task 1	60 US / L1 various SL English / 3 program levels	Task 1: Re-writing the 'Aluminium' passage /ANOVA on effect of level on measures	W/T (**).
Ho-Peng (1983) Task 2	60 US / L1 various SFL English / 3 program levels	Task 2: essays on one given topic, no time limit reported / ANOVA on effect of level on measures	W/T (**).
Larsen-freeman (1983) Study 2-task 1	109 US / L1 various, L2 English / 4 program levels	Task 1 re writing the aluminium passage. Task 2 / ANOVA on effect of level on measures	W/T (***).
Larsen-freeman (1983) Study 2-task 2	109 US / L1 various, L2 English / 4 program levels	Task 1 re writing the aluminium passage. Task 2 / ANOVA on effect of level on measures	W/T (**).
Homburg (1984), analysis 1	30 US / L1 various SFL English / 3 Holistic rating	30-minute essays on one of two given topics / ANOVA on effect of rating on measures	W/S, W/T, DC/T (**). T/S (X).
Homburg (1984), analysis 2	30 US / L1 various SFL English / 3 Holistic rating	30-minute essays on one of two given topics / Discriminant analysis correlating ratings & measures	W/S (*).
Bardovi-Harlig & Bofman (1989)	30 US / L1 various SFL English / 2 groups based on holistic judgment, same TOEFL score	45-minute essays on one of three topics / T tests comparing groups	C/T (X).
Linnarud (1986), analysis 1	63 juniors in high school / L1 Swedish, FL English / 2 groups SFL (Swedish) & NS. Holistic ratings of writing samples based 3-point holistic scale	40-minute description of a sequence of pictures / Analysis 1. T- tests comparing NS & SLS on measures	W/S (X).
Linnarud (1986), analysis 2	63 juniors in high school / L1 Swedish, FL English / 2 groups SFL (Swedish) & NS. Holistic ratings of writing samples based 3-point holistic scale	40-minute description of a sequence of pictures / Analysis 2. Correlations between holistic ratings and measures.	W/S (X).
Tedick (1990)	105 graduate students / L1 various SL English / three levels based on program placement	two essays, general and field-specific, 48-minute time limit for each / MANOVA on effect of program (and Topic) on measures	W/T (**).
Tomita (1990) analysis 1	258 high school students / L1 Japanese, FL English /3 school years (Level 1: Sophomores, Level 2: Juniors, Level 3: Seniors). 3 levels based on each school grade (low, mid, high)	3 written description of pictures, 5 minutes each / analysis 1: for each school year, correlation between grade and measures and ANOVA on effect of grade on measures	Level 2: Juniors W/T (*). Level 1: Sophomores & Level 3: Seniors W/T (X).
Tomita (1990) analysis 2	258 high school students / L1 Japanese, FL English / 3 school years, 3 levels	3 written description of pictures, 5 minutes each / Analysis 2 comparison of low	W/T (**).

	based on each school grade (low, mid, high)	and high holistic groups on measures	
Yau (1991)	60 high school students / L1 Chinese, FL English / 2 grade levels (9th, 13th grade) & NS	40-minute expository essay on one given topic / ANOVA on effect of group on measures	W/C, W/T (**).
Hirano (1991), analysis 1,2	158 US / L1 Japanese, FL English / 3 levels based on CELT scores	30-minute essays on one given topic, divided into 2 groups differing in audience / Correlations between CELT score & measures. Second analysis ANOVA on effect of level (and audience) on measures	W/T (**). W/C, C/T, DC/C. (*)
Hirano (1991), analysis 2	158 US / L1 Japanese, FL English / 3 levels based on CELT scores	30-minute essays on one given topic, divided into 2 groups differing in audience / Second analysis ANOVA on effect of level (and audience) on measures	W/T, C/T, DC/C (***). W/C (**).
Arnaud, (1992)	50 University Freshmen / French (L1), English (S/FL) / advanced learners, test scores on grammar and vocabulary tests	one-hour essays,6-8 weeks apart, on one given topic / correlations between test scores and measures on both compositions	W/T (*)
Kawata (1992),	44 school juniors / L1 Japanese, FL English / 4 levels based on school grades	50-minute essays on one given topic / ANOVA on effect of grade on measures	W/T (**). W/S (X).
Casanave 1994	16 university students / L1 Japanese / SFL English	journal entries from beginning, first third, and end of period, no time limit / no statistical tests, no group means, only an examination of individual patterns / selected students from two intact classes over a three-semester period, TOEFL score between 420-470	W/T, C/T, CT/T (*)
Ferris (1994)	160 entering to university / L1 Various, L2 English / 1-10 holistic rating, further divided into low level 60 students & advanced level 100 students	40 texts by each students / Discriminant analysis	Number of words, prepositional phrases, Negation, adverbials, 1s/2nd person & impersonal pronouns. (*)
Ishikawa, (1995) analysis group 2	57 US / L1 Japanese, FL English / 2 groups, intact classes, both beginning level	Two thirty-minute description of a picture sequence, 3 months apart / beginning class, 2 essays, 3 months apart / correlations between sums of individual	C/S (**). C/T, W/T, W/C, W/S, T/S (X).
Henry (1996)	67 US / L1 English, FL Russian / 4 school levels /	10-minute essay on the topic (me) / ANOVA on effect of level on measures	W/T (**).

Appendix 1. Table 2.

Studies on syntactic complexity after millennium (Studies are cross-sectional and non- CEFR studies on SC). In the table, *** denotes large effect size, ** moderate effect size, * small effect size, and (X) means non-significant results

Study	Writers / Participants / L1, L2, FL (languages) / # of levels / Proficiency	Samples / Analysis	(***) = Significant results with large effect size / (**) = Results with moderate effect size. / (*) = Results with small effect size. / (X) = Non-significant results
Neff et al. (2000)	105 (15 Professional authors, 30 in each fourth year, first year, & NS US) /L1 Spanish, FL English / 2 university years of non-natives, & professional NS, & University students natives	editorial articles (by NS 15), Argumentative by students / ANOVA on effect of years on measures	W/T, W/C, (Overall four groups) (***). C/TU (overall four groups) (*).
Beers & Nagy (2009)	41- 7th & 8th grade 13-14 years, Students were paid $15 for their participation /NS / middle school students / Woodcock–Johnson III-R (2001) tests	Expository texts conversational or narrative registers, partially written narrative and two different prompts of persuasive essays. 10 minutes to finalize the tasks / correlation	Positive correlation between Clause length & quality for essays, negative correlation between Clauses per T-unit & quality for essays positive correlation between Clauses per T-unit & quality for narratives. (*) No correlation between Clause length or T-unit length & quality for narratives, negative correlation between T-unit length and quality of narratives (X).
McNamara et al., (2010)	120 US / L1 English / 120 essays on 4different topics argumentative untimed & written outside of classroom essays. Referencing outside sources was allowed. Unequal number of students per prompt.	SAT rubric 1-6 / ANOVA	Significant differences between high vs. low proficiency on Left embeddedness. Non-significant differences between high vs. low proficiency on Number of words per sentence
Crossley et al., (2011)	120 /non-native speakers of English / Holistic rubric from the SAT essays of low vs. high	untimed argumentative essays college freshman / ANOVA	Significant differences for Modifiers per noun phrase across 9th grade, 11th grade and college freshman. significant differences for Left embeddedness between
Crossley, Weston, Sullivan, McNamara (2011)	202 /non- native English speakers / 9th, 11th grade, & college freshman, adolescents, and young adults	argumentative essays on SAT topics, 25 minutes essay, 62 essays from 62 9th-grade writers, 70 essays from 70 11th-grade writers, and 70 essays from 70 college freshmen / ANOVA,	Modifiers per noun across all grades (*)

		Discriminant Function Analysis	
Shih & Ma (2012)	657, L1 various, SFL English / advanced, upper-intermediate, intermediate, and the CEFR on GEPT, Grades	734 essays two prompts. / ANOVA	Significant results: CN/T between grades (7 vs. 11, 8 vs. 11); VPT between grades (7 vs. 10, 8 vs. 10, 9 vs. 10, 10 vs. 12, 7 vs 11, 8 vs. 11, 9 vs. 11) (**). MLS between 8 vs. 10; MLT between grades (7 vs. 10, 8 vs. 10, 9 vs. 10, 7 vs 11, 8 vs. 11, 9 vs. 11); MLC between grades (9 vs. 10, 9 vs. 11); C/T between grades (7 vs. 10, 8 vs. 10, 8 vs. 11); DC/C, DC/T & CT/T between grades (7 vs. 10, 8 vs. 10, 9 vs. 10, 7 vs 11, 8 vs. 11, 7 vs. 12, 8 vs. 12); CP/T & CP/C between grades (7 vs.10, 8 vs.7, 10 vs.7); CN/T between grades (7 vs.10, 8 vs. 10, 9 vs.11, 7 vs. 12, 8 vs. 12); CN/C between grades (7 vs. 11, 8 vs. 11, 9 vs. 11, 7 vs. 12) (*). C/S & T/S between all grades. MLS between grades (7 vs. 10, 9 vs. 10, 10 vs. 12, 7 vs 11, 8 vs. 11, 9 vs. 11, 7 vs. 12, 8 vs. 12, 8 vs. 7, 10 vs. 7); MLT between grades (10 vs. 12, 7 vs. 12, 8 vs. 12, 8 vs. 7, 10 vs. 7); MLC between grades (7 vs. 10, 8 vs. 10, 10 vs. 12, 7 vs 11, 8 vs. 11, 7 vs. 12, 8 vs. 12, 8 vs. 7, 10 vs. 7). C/T between grades (9 vs. 10, 10 vs. 12, 7 vs 11, 9 vs. 11, 7 vs. 12, 8 vs. 12, 8 vs. 7, 10 vs. 7); DC/C, DC/T & CT/T between grades (10 vs. 12, 9 vs. 11, 8 vs. 7, 10 vs. 7); CP/T & CP/C between the grades (8 vs. 10, 9 vs. 10, 10 vs. 12, 7 vs 11, 8 vs. 11, 9 vs. 11, 7 vs. 12, 8 vs. 12); CN/T between grades (8 vs. 10, 9 vs. 10, 8 vs. 7, 10 vs. 7); CN/C between grades (7 vs. 10, 8 vs. 10, 9 vs. 10, 10 vs. 12, 8 vs. 12, 8 vs. 7, 10 vs. 7; VP/T between grades (7 vs. 12, 8 vs. 12, 8 vs. 7, 10 vs. 7) (X)
Guo et al., (2013)	240, L1 various, EFL & ESL	20 minutes Integrated essay, 30 minutes argumentative essay Independent and integrated writing rubrics 1-5 proficiency. Correlations, Regression analysis	Significant results of integrated essays in training set of regression analysis: Past participle verbs, Verbs in base form, Personal pronouns, Verbs in 3rd person singular present form, Verbs in past tense, Modifiers per noun phrase, To infinitive, prepositional phrases. Significant results of independent essays in training set of regression analysis: Past Participle verbs, Verbs in non-3rd person singular present form, embeddedness clauses, Number of modifiers per

			noun phrase, personal pronouns, and verbs in base form.
Ai & Lu (2013)	L1 Chinese, EFL, NNS vs. NS, NNS-L vs. NS, NNS-H vs. NS, NNS-L, NNS-H /three student groups: (1) first- and second-year English major students (100 each) (NNS-low); (2) third- and fourth-year English major students (100 each) (NNS-high); and (3) American university students in (NS)	600 topics NS & NNS (214) wrote on 27 topics 26 topics argumentative & expository, timed, and untimed ANOVA and T-test	MLS, MLT & CN/T between (NNS vs. NS, NNS-L vs. NS, NNS-H vs. NS, NNS-L vs. NNS-H); MLC, DC/C, DC/T, CP/T & CNC between (NNS vs. NS, NNS-L vs. NS, NNS-H vs. NS) (*). MLC, MLS, MLT, DC/C, DC/T, CP/C, CP/T, T/S, CN/C between (NNS-L vs. NNS-H) CP/C, T/S between (NNS vs. NS, NNS-L vs. NS, NNS-H vs. NS) (X).
McNamara et al., (2013)	313 US /L1 English /313 timed 25 minutes Persuasive essays on two SAT test prompts	SAT rubric score between 1-6, correlations, regression analysis	Correlation between holistic ratings & High-Level constituents per word, Modifiers per noun phrase, Incidence of prepositional phrases, Incidence of S-bars, Incidence of verb phrases, Incidence of declarative sentences.
Bulté & Housen (2014)	45 L1 different ESL, holistic rating, gauging different aspects of L2 complexity (three rating scores: (i) the mean total score of all five rating scales and the scores of the scales, (ii) Language Use and (iii) Vocabulary)	Two samples of essays written 4 months apart, paired-Sample T-tests with Cohen's d effect size, linear regression	Correlation between holistic ratings over time, and complexity measure: MLS, MLTU, Overall Writing Quality. Correlations between SC measures MLT, MLS and Overall Holistic ratings. (*)
Youn (2014) cross-sectional	40 / L1 various, SL English / 3 levels, low intermediate university prepatory ESL program (TOEFL score 33 to 72), second group intermediate and high-intermediate (73 to 100), third group advanced proficiency (above 100 TOEFL score) or exited form EAP program	4 different tasks / correlation	C/TU overall across three groups (***). MLT, MLC, CTU. (*)
Crossley Allen, Kyle, McNamara al., (2014)	480 /L1 various, SL English / TOEFL IBT independent writing rubric 1-5	480 independent essays/ ANOVA Post-hoc	Significant correlations between human ratings and, noun phrase density, modifiers per noun phrase, agentless passives, perfect verb forms, incidence of conjuncts, determiners, minimal edit distance, adjective phrases
Crossley, Allen,	126 / L1 Native speakers of English /	SAT prompt 25 minutes timed essays/ SiNLP	significant correlations between sentence length, incidence of

Kyle, McNamara (2014)	11th grade, SAT rubric score 1-6		conjuncts, density of negations, second-person pronouns, part of speech: pronouns, number of words before the main verb, part of speech: third-person pronouns, part of speech: first-person pronouns & essay scores
Lu & Ai (2015) Cross-sectional	1600 essays written by intermediate & Advanced, L1 various, FL English & Natives / 20 random samples from/ at each 200 learners were rated on the CEFR levels /	essays on various topics (Argumentative essays) timed & untimed essays / T-tests	Significant results: Between NS & all NNS (MLC, CN/C, CN/C). Between NS vs. Japanese (MLS, MLC, MLT, C/T, CT/T, DC/C, DC/T, CP/C, CP/T, CN/C, CN/T, VP/T, C/S); Between NS vs. Chinese (MLS, MLT, C/T, CT/T, DC/C,DC/T,CP/T, T/S, VP/T, C/S); Between NS vs. Russian (MLC, MLT, CT/T, DC/C, DC/T, T/S, VP/T); Between NS vs. Tswana (MLS, MLC, C/T, CT/T, DC/C, DC/T, CP/C, T/S, CN/C, CN/T, VP/T, C/S); Between NS vs. French (T/S); Between NS vs. Bulgarian (CP/C, CP/T); Between NS vs. Japanese (MLS, MLC, MLT, CP/T, T/S, VP/T, C/S) (*) Non-significant results: Between Ns & all NNS (MLS, MLT, C/T, CT/T, DC/T, DC/C, CP/C, CP/T, T/S, VP/T, C/S); Between NS vs. Japanese (T/S); Between NS vs. Chinese (MLC, CP/C, CN/C, CN/T); Between NS vs. Russian (MLS, C/T,CP/C,CP/T,C/S); Between NS vs. Tswana (MLT, CP/T); Between NS vs. French (MLS, MLC, MLT, C/T, CT/T, DC/C, DC/T, CP/C, CP/T, CN/C, CN/T, VP/T, C/S); Between NS vs. Bulgarian ((MLS, MLC, MLT, C/T, CT/T, DC/C, DC/T, T/S, CN/C, CN/T, VP/T, C/S); Between NS vs. German (C/T, CT/T, DC/C, DC/T, CP/C, CN/C, CN/T) (X)
Crossley et al., (2015)	997 learners from different schools / not mentioned but appears “natives” / learners from 9th grade, 10th grade, 12th grade, and college freshman from 5 different geographical regions USA were double rated on standardized SAT rubric	9 persuasive prompts in essays, 25 minutes time limit Principal component analysis, training set of 673 essays and a test set of 304 essays, factor analysis, correlation against human essay scores	Correlation with essay quality of syntactic similarity across sentences, Average sentence length, Syntactic similarity across paragraphs, (**) Incidence verb phrases, Density verb phrases, Incidence of infinitives, Incidence of simple sentences, All verb incidence, Incidence of preposition phrases, Incidence of prepositions, Incidence of verb base forms, Words before main verb. (*) Syntactic Simplicity (X).

Crossley, Allen, & snow (2015)	87 / Native American school students aged average 15.6 years old with average grade 10.4, 30 identified English language learners / SAT rubric 1-6 scale	171 pre & post-test essays, Students attended 10 sessions (1 session/day) over a 2–4-week period. Participants wrote a pre-test essay during the first session and a post-test essay during the last session. The essays were written on two prompts (on the value of competition and on the role of image) counterbalanced across the pre-test and post-test essays regression analysis / correlation	significant correlations between Incidence of infinitives & human scores of essay writing quality
Mancilla et al., (2015)	102, NNSs & 141 NS US / L1 various, SL English / 486 discussion board postings	Texts from 102 High and low NNSs based on TOEFL scores, and 141 NS / ANOVA	DCC, DC/T, CP/C CNC between (NNS vs. NS), DC/C & DC/T between (NNS-L vs. NS). (*) MLC, MLS, MLT, CP/T, T/S, CN/T between the groups (NNS vs. NS), (NNS-H vs. NS), (NNS-L vs. NS), (NNS-L vs. NSS-H); DCC, & DC/T, between (NNS-H vs. NS & NNS-L vs. NNS-H) (X).
Wang & Slater (2016)	38 EFL learners, 15 English proficient learners / EFL learners (non-English majors, second-year university students versus English proficient users / To determine differences in syntactic complexity between Chinese EFL writers and online examples of personal statements	38 personal statements written by Chinese EFL students, and 15 personal statements examples taken from the Internet / T-tests	Sentence length, clause length, complex nominal per clause, complex nominal per T-unit. (*). T-unit length, verb phrases per T-unit, clauses per sentence, clauses per T-unit, dependent clauses per clause, dependent clauses per T-unit, complex T-units per T-unit, T-unit per sentence, coordinate phrases per T-unit, coordinate phrases per clause (X).
Tabard (2017)	90, 19-23 years US /L1 Persian EFL / argumentative task	Standardised English proficiency test three classes. Test was designed by researcher not approved by ETS or British council / ANOVA	Significant differences across three planning types of the proportion of Clauses to T-units, total # of verb forms. (*). No significant differences across three groups (classes) of learners on standardised English proficiency test (X).
Thilagha Jagaiah (2017)	Total 1029 students. 115 students-band 1-2; 914 sts-3, 4, 5 band) 8th grade school learners / L1 various, SL English / six to thirty-six using an Automated Essay Scoring (AES) engine	60 min each argumentative, informative/explanatory, and narrative essay (1029) of 8th grade. / Factor, regression analysis	ALL WORDS, ADV, DRPVAL, DRNP, DESSL, Temporal Connectives Incidence, All Connectives Incidence, Causal Connectives Incidence, number of modifiers per noun phrase, Minimal Edit Distance, Part of Speech, Minimal Edit Distance all words, Minimal Edit Distance, Lemmas,

	called Project Essay Grade (PEG; Page, 1966, 1994, 2003) & six sub-scores based on, the sum of six sub-scores was writing quality. (The students received six sub-scores ranging from one to six related to writing quality: overall development, organization, support, sentence structure, word choice, and mechanics using both holistic and Traits scores, which were like scores assigned by human raters (Chung & O'Neil, 1997; Shermis & Burstein, 2003). At-risk (band 1, 2 basic & below basic) & Not-at-risk (band 3-4-5 proficient, advanced learners)		between proficiency at-risk vs. not-at-risk. (*).
Yazdani (2018)	45 /L1 Iranian, SL English / Cambridge placement Test, Elementary, intermediate, advanced	essay on a topic/ correlation, ANOVA	MLC, MLT (elementary vs. advanced); MLT between (Intermediate vs. advanced) (***). MLS, MLC (Intermediate vs. advanced) (**). MLC, MLT between elementary vs. intermediate (X).
Kyle (2018)	240 / L1 various, SL English /data set was taken from the essays written for TOEFL test. 5 point based holistic rating	two prompts 240 each argumentative / Correlation, regression model	Across TOEFL scores significant for mean length of clause (*). Coordinate phrases per clause (X).
Jiang et al., (2019)	410 / L1 Chinese, EFL English / four writing proficiency levels written by beginner and intermediate L2 English learners & The rating rubric is adapted from The Preliminary English Test of Cambridge (PET)	30 minutes-410 narratives / ANOVA, MANOVA	MLC MLT MLS DC/C CN/C between proficiency level 2 and 3. MLT MLS DC/C between 1 & 2. MLC MLT between 3 & 4. (*). MLC between 1 & 2, 2 & 3. MLS DC/C CN/C between 3 & 4 (X).
Casal & lee (2019)	280 ESL undergraduate / L1 various, ESL / High, mid, and low, writing quality	research papers on range of topics / MANOVA (L2SCA linguistic analysis)	MLT, MLC, CN/C between high vs low, CN/C between mid vs. low. C/T & T/S between High vs. mid, high vs. low, mid vs. low. CN/C &

	(operationalized as instructor assigned grades)		MLT High vs. Mid, Low vs. Mid. MLC between High vs. Mid.
Shadloo et al., (2019)	104-EFL US & Schools, 18-30 years old / L1 Persian, EFL learners / Low, mid, high	argumentative essays / ANOVA	Across low, mid, high significant results on Mean length of T-unit (MLT), Mean length of clause (MLC), Verb phrases per T-unit (VP/T), Complex nominal per T-unit (CN/T), Complex nominal per clause (CN/C). (*). Across low, mid, high non-significant results on Clauses per sentence (C/S), Clauses per T-unit (C/T), Dependent clauses per clause (DP/C), Dependent clauses per T-unit (DP/T), T-unit per sentence (T/S), Complex T-unit ratio, Coordinate phrases per T-unit (CP/T), Coordinate phrases per clause (CP/C), (X).

Appendix 1. Table 3.

Syntactic complexity features that significantly separated the CEFR levels in adult EFL learners' writing in previous research

CEFR levels		Syntactic complexity features
C1	vs	*Adult EFL learners*
C2		Sentence length (Hawkins and Filipović 2012), noun phrase incidence, modifiers per noun, sentence syntax similarity (Green 2012). Modifiers per noun phrase between C2 vs. below C2 level Banerjee et al., (2015). Clauses per T-unit, dependent clauses per T-unit, dependent clauses per clause, mean length of clause, verb phrases per T-unit, complex nominals per T-unit, and complex nominals per clause between C1 versus C2. Paquot (2019).
B2	vs	*Adult EFL learners*
C1		Sentence length (Hawkins and Filipović 2012) noun phrase incidence, modifiers per noun, sentence syntax similarity (Green 2012). Clauses per T-unit, dependent clauses per T-unit, dependent clauses per clause, mean length of clause, verb phrases per T-unit, complex nominals per T-unit, and complex nominals per clause between B2 versus C1, Paquot (2019).
B1 vs B2		*Adult EFL learners*
		Sentence, clause and T-unit length, clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, dependent clauses per T-unit, coordinate phrases per clause, coordinate phrase per T-unit, T-units per sentence, complex nominal per clause, complex nominal per T-unit, verb phrases per T-unit, clauses per sentence (Polat et al., 2020). Sentence length (Hawkins and Filipović 2012), clauses and dependent clauses per T-unit; dependent clauses per clause (Kim 2004). Mean sentence length, clauses per sentence and dependent clauses per sentence learners at B1.2 level between EFL and ESL (Qi 2014). T-unit length, clause length and complex nominal per clause; T-unit length and clause length; complex nominal per clause between B1.1 vs. B1.2; B1.2 vs. B2 (Yoon (2017)
A2	vs	*Adult EFL learners*
B1		Sentence, clause and T-unit length, clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, dependent clauses per T-unit, coordinate phrases per clause, coordinate phrase per T-unit T-units per sentence, complex nominal per clause, complex nominal per

	T-unit, verb phrases per T-unit, clauses per sentence (Polat et al., 2020). T-unit length, clause length and complex nominal per clause; T-unit length and clause length; complex nominal per clause between A2 vs. B1.1; (Yoon (2017). Sentence length (Hawkins and Filipović 2012), sentence length; length of clause; subordinate clauses per T-unit (Alexopoulou et al., 2017) compound, and complex sentence ratios; coordinate and dependent clause ratios; noun phrases per clause (Kim 2004) *Young EFL learners* T-unit length (Gyllstad et al., 2014, Verspoor et al., 2012); length of clauses. Clauses per T-unit (Gyllstad et al., 2014). Sentence length (Lahuerta Martínez 2018)
A1vs	*Adult EFL learners*
A2	Sentence length; length of clause; subordinate clauses per T-unit (Alexopoulou et al., 2017) *Young EFL learners* T-unit length (Verspoor et al., 2012)

Appendix 1. Table 4.

A summary of syntactic structures counted by L2SCA, table syntactic structures by Lu (2010, pp. 7–13).

Structure	Description	Examples
Word	A sequence of letters that are bounded by white space	I ate
Verb phrase		ate pizza, was hungry
Complex nominal	i. nouns with modifiers ii. ii. nominal clauses iii. iii. gerunds and infinitives that iv. function as subjects	v. red car vi. ii. I know that she is hungry. vii. iii. Running is invigorating
Coordinate phrase	Adjective, adverb, noun, and verb phrases connected by a coordinating conjunction	She eats pizza and smiles
Clause	A syntactic structure with a subject and a finite verb	I ate pizza because I was hungry I ate pizza because I was hungry
Dependent clause	A finite clause that is a nominal, adverbial, or adjective clause	I ate pizza because I was hungry
T-unit	An independent clause and any clauses dependent on it	I ate pizza I ate pizza because I was hungry
Complex T-unit	A T-unit that includes a dependent clause	I ate pizza because I was hungry
Sentence	A group of words bounded by sentence-ending punctuation (., ?, !, ”, …)	I went running today.

Note: Adapted from Lu (2010, pp. 7–13).

APPENDIX 2.
The rating scale used in the DIALUKI project and in the dissertation (combination of seven CEFR scales)

	OVERALL WRITTEN PRODUCTION	WRITTEN INTERACTION	CORRESPONDENCE & NOTES, MESSAGES, FORMS	CREATIVE WRITING & THEMATIC DEVELOPMENT & COHERENCE AND COHESION
A1	Can write simple isolated phrases and sentences.	Can ask for or pass on personal details in written form.	Can write a short simple postcard. Can write numbers and dates, own name, nationality, address, age, date of birth or arrival in the country, etc. such as on a hotel registration form.	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do. Can link words or groups of words with very basic linear connectors like 'and' or 'then'.
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.	Can write short, simple formulaic notes relating to matters in areas of immediate need.	Can write very simple personal letters expressing thanks and apology. Can take a short, simple message provided he/she can ask for repetition and reformulation. Can write short, simple notes and messages relating to matters in areas of immediate need.	Can write about everyday aspects of his/her environment, e.g. people, places, a job or study experience in linked sentences. Can write very short, basic descriptions of events, past activities and personal experiences. Can write a series of simple phrases and sentences about their family, living conditions, educational background, present or most recent job. Can write short, simple imaginary biographies and simple poems about people. Can tell a story or describe something in a simple list of points. Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points. Can link groups of words with simple connectors like 'and', 'but' and 'because'.
B1	Can write straightforward connected texts on a range of	Can convey information and ideas on abstract as well as concrete topics, check information and	Can write personal letters giving news and expressing thoughts about abstract or cultural topics such as music, films.	Can write straightforward, detailed descriptions on a range of familiar subjects within his/her field of interest. Can write accounts of experiences, describing feelings and reactions in simple connected text.

	familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence.	ask about or explain problems with reasonable precision. Can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point he/she feels to be important.	Can write personal letters describing experiences, feelings and events in some detail. Can write notes conveying simple information of immediate relevance to friends, service people, teachers and others who feature in his/her everyday life, getting across comprehensibly the points he/she feels are important. Can take messages communicating enquiries, explaining problems.	Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points. Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.

	OVERALL WRITTEN PRODUCTION	**WRITTEN INTERACTION**	**CORRESPONDENCE & NOTES, MESSAGES, FORMS**	**CREATIVE WRITING & THEMATIC DEVELOPMENT & COHERENCE AND COHESION**
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.	Can express news and views effectively in writing, and relate to those of others.	Can write letters conveying degrees of emotion and highlighting the personal significance of events and experiences and commenting on the correspondent's news and views.	Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play. Can develop a clear description or narrative, expanding and supporting his/her main points with relevant supporting detail and examples. Can use a variety of linking words efficiently to mark clearly the relationships between ideas.

				Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse though there may be some 'jumpiness' in long contribution.
C1	Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.	Can express him/herself with clarity and precision, relating to the addressee flexibly and effectively.	Can express him/herself with clarity and precision in personal correspondence, using language flexibly and effectively, including emotional, allusive and joking usage.	Can write clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind. Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion. Can produce clear, smoothly flowing, well-structured text showing controlled use of organisational patterns, connectors, and cohesive ideas.
C2	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.	As C1	As C1	Can write clear, smoothly flowing, and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted. Can create coherent and cohesive text making full and appropriate use of a variety of organisational patterns and a wide range of cohesive devices.

Appendix 3

Examples of Sindhi and Finnish EFL learners' texts and the values of selected syntactic complexity indices related to them (from the L2 Syntactic Complexity Analyzer) based on the Mobile Phone task

A1 level samples of Sindhi EFL learners' texts

1.

No mobile phones at School.

I am faver in this matters. No mobile phone at School time because this for study. We are come to the school for purpose and when we come to school with cell phone. When we use it, this is absolute wrong then, and this good and stick rule of all School. Student and all students from school or college.

2.

Mobile phone have only use of contact and other side. Study needs concentrate and if a boy use mobile phone. So, how can he give the time to education? Mobile phone stealing make problem.
Mobile phone ringing can disturb the students. Mobile phone can annoyed the student and there is no any kind of advantage or it will not give any type information to boys.

3.

No mobile in school.
Mobile must not be allowed in schools because boys will play games, audios, videos, applications and a lot of other graphics. So, that all treatments will not allow for studies. All the students will situate, teacher will in classroom. Teacher is reading and mobile is ringing.

Selected syntactic complexity indices for the Sindhi A1 samples:

sample	nbr of words	MLS	MLT	MLC	C_S	VP_T	C_T
1	72	12	18	14.4	0.83	1.50	1.25
2	72	12	10.29	8	1.17	1.14	1.00
3	60	12	12	7.5	1.40	1.40	1.40

sample	DC_C	DC_T	T_S	CT_T	CP_T	CP_C	CN_T	CN_C
1	0,028	0,035	0.67	0,035	0,059	0,042	1.75	1.40
2	0,000	0,000	1.17	0,000	0,020	0,020	0.71	0.71
3	0,010	0,014	1.00	0,014	0,014	0,010	1.00	0.71

A1 level samples of Finnish EFL learners' texts

1.

No mobile phones at school!

I think that mobiles not makes good in us. You can't hear teacher same than you send sms in your friends. Mobile phones give so much attention in lesson.

2.

I think that mobile phones doesn't belong
to school. They always disturb (häiritsevät) lessons.

3.

No mobile phones at school!
Mobile phones are too noisy for school and they are annoying for teachers.

Selected syntactic complexity indices for the Finnish A1 samples:

sample	nbr of words	MLS	MLT	MLC	C_S	VP_T	C_T
1	35	8.75	11.67	7	0,059	1.67	1.67
2	14	8.50	8.50	8.50	1.00	1.50	1.00
3	18	9.00	9.00	6.00	1.50	1.00	1.50

sample	DC_C	DC_T	T_S	CT_T	CP_T	CP_C	CN_T	CN_C
1	-	0.67	0.75	0.67	0	0	1.67	1
2	0.50	0.50	1.00	0.50	0	0	2.00	2
3	0	0	1.00	0	0	0	0.50	0.33

A2 level samples of Sindhi EFL learners' texts

1.

No mobile phones at school.
Nowadays, there are a lot of mobiles which are used at home. In home, if there are five members of family. Everybody has his own mobile but not only at home but when a student go to school, he also carry there.
First of all, student must not use mobile. For what work, he use mobile. There is nothing to use mobile for students.
There are some reason to not use mobile at school. When mobile rings at School, the lecture of teacher will be disturbed. When student use mobile at school, I can not learn more because his full mind will be at mobile. But some students who are out door of city and in that condition, he most use but not at school. He can use mobile at hostel or at the time of holiday.

2.

Some things, such like mobiles, they are playing a great role in disaster of students' studies.
There are students mostly busy in the reading of messages. They used to listen songs, watch videos, etc due to waste the time. They mostly purchase and think every time about mobiles. They are fond of showing off by having a expensive mobiles. They use internet in the mobile but in wrong way mostly Facebook. Due to mobile their, interest in studies are totally finished. According to health, it effect our eyes and losses our mind and other thing. We always look down there is a Newton force on our neck. So, with the passage of time, it remains bend.

3.

No mobile Phones at school.
Mobile is a modern technology of science but it is not for to use in school. Again, if we are them then certain problems are found, which are given below.
Mobile phones is a cause for disturbance at school. It give a bad impression to the citizens of a city of that particular school. Complete atmosphere, it is just considered as tobacco in the hand at school.

Selected syntactic complexity indices for the Sindhi A2 samples:

sample	nbr of words	MLS	MLT	MLC	C_S	VP_T	C_T
1	142	11.83	11.83	10.14	1.17	1.50	0,053
2	117	11.70	11.70	11.70	1.00	1.70	0,042
3	72	12.00	12.00	9.00	1.33	1.50	0,065

sample	DC_C	DC_T	T_S	CT_T	CP_T	CP_C	CN_T	CN_C
1	0,020	0,023	1.00	0,012	0,000	0,000	0,071	0,056
2	0,000	0,000	1.00	0,000	0,014	0,014	0,063	0,063
3	0,017	0,023	1.00	0,012	0,000	0,000	0,083	0,076

A2 level samples of Finnish EFL learners' texts

1.

No mobile phones at school!
I think that are stupid idea.
What is so bad all phones? If it ring couple times in class the teacher take it. It's wrong, because teacher can't take the student phone because it he phone isn't are the teacher's. Many listening music at school with phone.
If "No mobile phones at school!" come true, all student gonna do bad things.

2.

Mobile phone is very important at school because every one needs that.
You not survive without mobile phone.
Example: your mom calling you and says: "I go store, I am not go straight home, do you have keys?" If you don't have phone you don't know that mom don't go straight home.

3.

No mobile phones at school!
That's not right. My opinion is that everybody could keep mobile phones on school.
It's true that mobile phones aren't allowed to use at lessons but in breaks everybody need them. Sometimes your parents could have very important thing to speak and if mobile phones aren't allowed in school you can't hear it.
At lessons mobile phones should be shut down, but everybody have to be chance call or sent text message on free time.

Selected syntactic complexity indices for the Finnish A2 samples:

sample	nbr of words	MLS	MLT	MLC	C_S	VP_T	C_T
1	65	8.75	11.67	6.36	1.38	1.83	1.83
2	52	13.75	13.75	5.50	2.50	2.50	2.50
3	80	14.17	10.63	8.50	1.67	1.75	1.25

sample	DC_C	DC_T	T_S	CT_T	CP_T	CP_C	CN_T	CN_C
1	0.45	0.83	0.75	0.50	0	0	1.50	0.82
2	0.40	1.00	1.00	0.75	0.25	0.10	1.50	0.60
3	0.30	0.38	1.33	0.50	0.13	0.10	1.13	0.90

B1 level samples of Sindhi EFL learners' texts

1.

No Mobile phones at school.
Actually, all things depend upon a men's own thoughts. If we say that mobile phones should be prohibited then it also could be said 'good' or if some one say that it is 'bad'. They could be right because there are many benefits of keeping mobiles phones along with.
Benefits, We can get information arising twenty hours in our mind.
Since, it is communication way so we would know all the Information around the world.

We will be able to connect ourself with new era of information technology and the new discoveries, inventions and information, news about all around us.
Harms, it disturbs the students in studying. since there are also miss uses of it like, films, movies, games, miss using of messaging. If used in negative way or wasting of time ordered.
Extra use of mobile losses the activeness of mind. As a report has shown and proved this.

2.

No mobile phones at school.
Mobile phone is very helping device all over the world. We can easily send a message to anyone. It has been the fastest tool of convey . But after all, it has been dangerous for those who have misuse of it.
Nowadays, many crimes have been committed due to mobile phones. Even, many girls have escaped. Many bombs were blasted and even are blasted through mobile phone. And, it has been proved that mobile phone is heavy enemy of students because it does not let the students study with full concentration. It drives out the attention of students from study and there is huge need of strive hard. Number second, the students mostly misuse of it and they are gone to sexually, and it direct attack on the mind of students so students must avoid of misuse of mobile when they use it.

3.

No mobile phones at School.
Today, the modern world, Science has made many electrical things which are very useful for our daily life like, televisions, mobile phones, radio, Laptops, tablets, computers and many other. In the school, mobile phone must not be our there. Do not allow to the Students to take mobile phones in the class. First reason is that there will be disturbance in environment. Second train of reading and teaching will break down and teacher can not teach them properly. It is duty of a teacher that take care of such these conditions. So, that teacher and student can not give and take education. Mobile phones are harmful for students but any student have a serious problem. He must get but do not disturb the environment .

Selected syntactic complexity indices for the Sindhi B1 samples:

sample	nbr of words	MLS	MLT	MLC	C_S	VP_T	C_T
1	160	13.3	16	8.889	1.50	0,090	1.80
2	147	13.4	12.25	8.647	1.55	0,082	1.42
3	129	12.9	10.75	9.214	1.40	0,065	1.17

sample	DC_C	DC_T	T_S	CT_T	CP_T	CP_C	CN_T	CN_C
1	0,031	0.80	0.83	0.40	0.40	0,015	1.50	0.83
2	0,020	0.42	1.09	0.42	0.17	0,008	1.75	1.24
3	0,015	0.25	1.20	0.25	0.33	0,020	1.33	1.14

B1 level samples of Finnish EFL learners' texts

1.

No mobile phones to school! Umm what is really point of this? I mean seriously. Why would anyone say that you can't take your mobile phone with you?
What about you break your leg and you don't have mobile phone with you. So, mobile phones, you can take it to school.

2.

No mobile phones at school!
Mobile phones can disturb classes in school if they are used in lessons. Mobile phones should be turned off in lessons and used only in lunch breaks. Many pupils need mobile phone after school to call, for example, ride to home. So if mobile phone is used smartly it is okay at school.

3.

No mobile phones at school.
I have to disagree with that. If someone has to call a ride to home after school for example and wouldn't have a mobile phone what would that person do?
It's true that mobile phones sometimes erupts lesson, but they are pretty useful.
If you have to inform something to someone quickly, you can text or call really fast.
Of course in lessons mobile phones should be closed so they wouldn't interrupt other peoples learning.

Selected syntactic complexity indices for the Finnish B1 samples:

sample	nbr of words	MLS	MLT	MLC	C_S	VP_T	C_T
1	51	8.83	10.60	6.63	1.33	1.60	1.60
2	58	11.60	14.50	9.67	1.20	2.00	1.50
3	80	13.83	13.83	8.30	1.67	2.33	1.67

sample	DC_C	DC_T	T_S	CT_T	CP_T	CP_C	CN_T	CN_C
1	0.38	0.60	0.83	0.40	0	0	1.40	0.88
2	0.33	0.50	0.80	0.50	0.25	0.17	1.00	0.67
3	0.40	0.67	1.00	0.67	0.33	0.20	1.33	0.80

B2 level samples of Finnish EFL learners' texts

1.

No mobile phones at school!
It's very irritating to hear that "beep beep" or "ring ring" when you are trying to concentrate to difficult tasks. So in my opinion mobile phones shouldn't be allowed at school. I don't even understand why they even are allowed in first place. No one benefits from them during a school time. I would highly appreciate a new rule concerning mobiles!

2.

No mobile phones at school. I think the rule no mobile phones at school is good when you are in the class.

When teacher is teaching it's not polite to speak to phone or send messages to friends. At breaks using phones should be allowed, because then people do not disturb anyone. If mobile phones are not allowed at all, everyone still uses them to be rebellious. It is okay to call someone to pick you up after school or call to dentist at breaks.

3.

No mobile phones at school! It is not unusual that during the lesson you can hear many bleeping sounds followed by Sorry!. I understand that friends are an important part of people's lives, but is it really necessary to be in touch with them all the time?

In school we learn not only the subjects we study, but to respect and live with other people. I do not mind if someone talks on his phone during the recess, but when the lesson starts should all phones be turned off.

Selected syntactic complexity indices for the Finnish B2 samples:

sample	nbr of words	MLS	MLT	MLC	C_S	VP_T	C_T
1	66	11.50	13.80	9.86	1.17	1.80	1.40
2	86	14.33	17.20	8.60	1.67	3.00	2.00
3	90	15.00	18.00	10.00	1.50	2.00	1.80

sample	DC_C	DC_T	T_S	CT_T	CP_T	CP_C	CN_T	CN_C
1	0.29	0.40	0.83	0.40	0.20	0.14	1.00	0.71
2	0.50	1.00	0.83	0.80	0.40	0.20	1.00	0.50
3	0.44	0.80	0.83	0.60	0.20	0.11	2.20	1.22

ORIGINAL PAPERS

I

INVESTIGATING SYNTACTIC COMPLEXITY IN EFL LEARNERS' WRITING ACROSS COMMON EUROPEAN FRAMEWORK OF REFERENCE LEVELS A1, A2, AND B1.

by

Khushik, G. A., & Huhta, A. 2020.

Applied Linguistics, 41(4), 506-532.

https://doi.org/10.1093/applin/amy064

Reproduced with kind permission by Oxford University Press.

Investigating Syntactic Complexity in EFL learners' writing across Common European Framework of Reference Levels A1, A2, and B1

Abstract

The study investigates the linguistic basis of CEFR levels in English as a foreign language learners' writing. Specifically, it examines whether CEFR levels can be distinguished with reference to syntactic complexity (SC) and whether the results differ between two groups of EFL learners with different first languages (Sindhi and Finnish). This sheds light on the linguistic comparability of the CEFR levels across L1 groups. Informants were teenagers from Pakistan (N=868) and Finland (N=287) who wrote the same argumentative essay that was rated on a CEFR-based scale. The essays were analysed for 28 SC indices with the L2 Syntactic Complexity Analyzer and Coh-Metrix. Most indices were found to distinguish CEFR levels A1, A2 and B1 in both language groups: the clearest separators were the length of production units, subordination, and phrasal density indices. The learner groups differed most in the length measures and phrasal density when their CEFR level was controlled for. However, some indices remained the same, and the A1 level was more similar than A2 and B2 in terms of SC across the two groups.

INTRODUCTION

This study investigates the linguistic basis of the Common European Framework of Reference (CEFR; Council of Europe 2001) by focusing on syntactic complexity (henceforth SC). The CEFR has become increasingly important in foreign and second language (L2) education particularly in Europe (Hulstijn *et al.* 2010) but also beyond. Besides providing rich descriptions of learning and using languages, the CEFR includes scales defining what language learners can do in an L2 at different stages (levels) of proficiency. These levels can also be understood as general descriptions of stages in L2 development. As will be elaborated below, the CEFR scales do not define syntactic complexity or other linguistic concepts in detail, neither are they based on solid empirical research on L2 learning.

The current study is exploratory as it is not based on specific hypotheses about which aspects of SC might characterise particular CEFR levels or distinguish between them. Rather we investigate a wide range of indices used in previous research. Besides the breadth of SC indices covered, another important feature of the study is that it focuses on two first language (L1) groups, Sindhi and Finnish, learning the same foreign language (English), which allows us to examine the linguistic comparability of the CEFR levels across two different L1 groups languages.

Defining syntactic complexity

Complexity and complex systems have been studied extensively in different fields ranging from natural sciences to social sciences and, for the past two decades, also in SLA and L2 writing research (Wolfe-Quintero *et al.* 1998; Ortega 2003; Bulté and Housen 2014). However, there is no consensus on the definition of complexity apart from the recognition that it is a very complex concept that comprises many levels and dimensions (Norris and Ortega 2009). Recently, Bulté and Housen (2012, 2014) have proposed a framework describing the different aspects of complexity and how complexity relates to difficulty. Building on theoretical discussions of complexity by, e.g., Dahl (2004), Kusters (2008) and Miestamo (2008), Bulté and Housen (2012) divide L2 complexity into relative and absolute thereby distinguishing difficulty from complexity. Difficulty relates to relative complexity: the amount of cognitive effort certain linguistic features require when used or acquired by L2 learners (see Housen and Simoens, 2016, for a discussion of difficulty). The effort varies between learners depending on their stage of L2 development, L1 background and motivation, which means different linguistic features are not equally difficult for all learners. Absolute complexity is defined in objective terms as the number of and connections between the different components of a linguistic feature. Absolute complexity can be further divided into linguistic, propositional and discourse-interactional complexity.

Syntactic complexity is part of linguistic complexity, and as far as individual linguistic features are concerned, the most relevant aspect of linguistic complexity is structure complexity, which can be divided into functional and formal types. According to Bulté and Housen (2012: 24), "[f]unctional complexity refers to the number of meanings and functions of a linguistic structure and to the degree of transparency, or multiplicity, of the mapping between the form and meanings/functions of a linguistic feature". Some structures have clear one-to-one mapping between meaning and form,

whereas others lack such straightforward mappings. Bulté and Housen (2012) mention the English plural marker (–s) as an example of the former and the English third person singular marker of the present tense (-s) to illustrate the latter. According to Bulté and Housen's analysis, formal complexity can be defined as the number of discrete components of the linguistic form or as the number of operations needed to turn a base structure into the target structure (e.g., from active to passive form).

Bulté and Housen (2012) argue that when investigating such aspects of linguistic complexity as syntactic complexity it is important to consider three levels of construct specification: theoretical, observational, and operational. The abstract theoretical level concerns the number of components that a linguistic structure comprises and how these components relate to each other (e.g., embeddedness). The operational level concerns the different manifestations of the forms in language use that contribute to sentential, clausal or phrasal complexity. The third, operational, level relates to the analytical measures that yield quantitative indices of complexity.

The current study adopts Bulté and Housen's (2014: 45–46) definition and considers complexity "as an absolute, objective, and essentially quantitative property of language units, features, and (sub) systems thereof in terms of (i) the number and the nature of discrete parts that the unit/feature/system consists of and (ii) the number and the nature of the interconnections between the parts".

CEFR scales

Investigations of the linguistic basis of the CEFR levels are needed because these levels are meant to be language-independent and describe how learners *use* a language, not which linguistic features characterise levels. However, understanding, e.g., how specific linguistic features might change between CEFR levels would assist us in evaluating the validity of the descriptions and in developing more level-appropriate teaching/learning materials, courses, and assessments (Hulstijn *et al.* 2010).

The CEFR contains some references to linguistic elements, even to SC, but they are unsystematic and ambiguous, and not linkable with particular levels. The clearest references to SC are found in the Overall written production scale (CoE 2001: 61), which mentions *simple phrases and sentences* at A1 and A2, and 'linking a series of shorter discrete elements into a linear sequence' at B1. A rare reference to SC occurs in the General linguistic range scale at B2 (CoE 2001: 110): "Has a sufficient range of language to be able to give clear descriptions, express viewpoints, and develop arguments ... using some complex sentence forms to do so". Most mentions of complexity in the CEFR refer to texts, topics, information, instructions, interactions or lines of argument, not syntax (Table 1 in Supplementary Data). Besides lacking linguistic detail, the CEFR scales have another shortcoming: they are not informed by theories of L2 development (CoE 2001: 21) or SLA research (Hulstijn *et al.* 2010).

Given these limitations, both the Council of Europe (CoE 2001) and scholars have called for research on the linguistic characteristics of the CEFR levels (e.g. Alderson 2007; Hulstijn 2007; Wiśniewski 2017). Researchers have responded (Bartning *et al.* 2010) and published on various aspects of vocabulary knowledge at CEFR levels such as vocabulary size (Milton 2013) and diversity (Treffers-Daller *et al.* 2016). Corresponding studies on syntax are described next in the literature review.

LITERATURE REVIEW

The relationship between SC and language proficiency has been examined extensively (e.g. Wolfe-Quintero *et al.* 1998; Ortega 2003; McNamara *et al.* 2010; Lu 2011; Guo *et al.* 2013; Kyle 2016). However, only some studies have operationalised language proficiency with reference to the CEFR; such studies focusing on EFL writing are reviewed below and summarised in Table 1.

An early study by Kim (2004) investigated CEFR-rated scripts from 33 Chinese EFL university students. Kim took clauses and T-units as the basis of analysis (T-unit is defined by Banerjee et al. (2007: 41) as "the unit generated when text is divided into the smallest possible independent segments, without leaving sentence fragments behind. Each T-unit consists of a main clause and all the subordinate clauses that belong to it"). Kim investigated three aspects of SC: (1) variety of structures (adverbial, adjective, and nominal clauses per clause), (2) number of subordinate clauses (clauses and dependent clauses per T-unit, dependent clauses per clause), and (3) shift from clauses to phrases (prepositional, participial, gerund, and infinitive phrases per clause). She found clear differences between A2 and B2 levels in all these measures except for nominal clauses per clause and gerund phrases per clause. Differences between A2 and B1 were not very clear but more pronounced between B1 and B2. Strong points in Kim's study include the direct rating of the scripts on the CEFR levels and the relatively wide range of SC indices examined. However, the study investigated a rather small group of learners who represented only one L1 background.

Studies conducted in the English Profile Programme on learners' performances on language test tasks, which forms the large-scale Cambridge Learner Corpus, have discovered that sentence length increases significantly between each adjacent level from A2 to C2 (Hawkins and Filipović 2012). Green (2012) reported significant differences in the noun phrase incidence and the number of modifiers per noun between B2 and C1. Green also found C1 and C2 to differ in terms of sentence syntax similarity. The advantage of the English Profile studies is that they cover almost the whole range of CEFR levels and are based on a very large learner corpus. The project has not investigated possible differences in SC due to learners' L1 background since the learners in their studies have very heterogeneous backgrounds (age, L1), and the coverage of SC indices has been limited. Furthermore, learners' placement on the CEFR levels is not done by rating them directly against CEFR-based scales but indirectly through their performance on examinations targeting specific levels.

Verspoor *et al.* (2012) investigated 437 young (aged 12-15) Dutch EFL learners who wrote one descriptive text on topics which varied depending on the learners' grade level. The scripts were rated on a 5-point scale corresponding to CEFR levels A1.1, A1.2, A2, B1.1, and B1.2. The authors found the mean T-unit length to increase across levels and significantly differentiate A1.2 vs B1.1, and A2 vs B1.2. They also reported the proportion of simple vs complex sentences to be a fairly good separator of levels, with the clearest leap taking place between A1.2 and A2. They further found the proportion of dependent clauses to be a particularly good separator and finite relative clauses to increase steadily across all levels but most clearly between A2 and B1.1. While Verspoor *et al*. rated their learners' texts directly on the CEFR levels and investigated a large number of learners; their study focused on only one L1 group and covered a limited range of SC indices.

Gyllstad *et al.* (2014) examined 54 Swedish EFL learners who wrote an email and a story. The three SC indices they investigated correlated significantly with the

rated CEFR levels: mean length of T-units (.48), mean length of clauses (.31), and clauses per T-unit (.46). The researchers divided the texts broadly into A and B levels on the CEFR and found all three indices to separate these two broad levels. Although Gyllstad et al. used direct CEFR ratings for the texts, they, too, investigated only one relatively small L1 group, used only a few SC indices and very broad CEFR scale categories.

Alexopoulou *et al.* (2017), using the EFCAMDAT, an open-access corpus (http://corpus.mml.cam.ac.uk/efcamdat), investigated SC indices in EFL writers' texts and found sentence length to increase across all CEFR levels. They also reported a clear increase in subclausal density (length of clause) from A2 to B2 and in subordination (number of subordinate clauses per T-unit) between each successive level from A1 to B2, but it is not clear if these changes were statistically significant. The study investigated the whole CEFR range by using a large dataset. However, it included only three SC indices and was based on learners with varied L1s. Furthermore, the relationship between the 16 proficiency levels in the corpus and the CEFR levels is uncertain.

Finally, Lahuerta Martínez (2018) investigated 188 secondary level Spanish EFL learners who wrote on the same topic requiring an expression of opinion. The students came from two grades that presumably represented A2 and B1 levels. The study found that sentence length, compound and complex sentence ratios, coordinate and dependent clause ratios, and noun phrases per clause separated the grade levels significantly. The study was fairly large-scale and all participants completed the same task under the same conditions. However, only one L1 group was investigated and their placement on the CEFR levels is uncertain as it was based on learners' grade levels.

Table 1: Previous studies on syntactic complexity in EFL writing across CEFR levels

Researchers	**Indices**	**CEFR levels that the indices separate**
Hawkins & Filipović (2012)	Sentence length	A2 vs B1, B1 vs B2 B2 vs C1, C1 vs C2
Green (2012)	Noun phrase incidence; number of modifiers per noun; sentence syntax similarity	B2 vs C1 C1 vs C2
Gyllstad et al. (2014)	T-unit length; clause length; clauses per T-unit	A2 vs B1
Verspoor et al. (2012)	T-unit length	A1 vs A2, A2 vs B1
Kim (2003)	Adverbial, adjective & nominal clauses per clause; clauses and dependent clauses per T-unit; dependent clauses per clause; prepositional, participial, gerund and infinitive phrases per clause	A2 vs B2 (more clearly between B1/B2 than between A2/B1)
Alexopoulou et al., (2017)	Sentence length; Mean length of clause; subordinate clauses per T-unit	A1 / A2 to B2
Lahuerta Martínez (2018)	Sentence length, compound and complex sentence ratios; coordinate and dependent clause ratios; noun phrases per clause	A2 vs B1

Since the present study differs from previous research in that it investigates two L1 groups of EFL learners, we complement the literature review with a scrutiny of studies that explicitly compare texts written by EFL learners with different first languages.

Apparently, the only CEFR-related study has been by Lu and Ai (2015) who used international corpora to compare college level EFL learners representing several L1 groups (N=200 per group) with native English-speaking university students who all wrote argumentative essays. The design of the study and the CEFR level distributions (none of the L1 groups represented only one level) make conclusions tentative but their results suggested that certain L1 groups differed in terms of SC at B2 and C1 levels. For example, at B2, speakers of Japanese and Chinese differed from Tswana (from the Niger-Congo language family) speakers in sentence and T-unit length, and particularly in clauses per T-unit, complex T-units per T-unit, and dependent clauses per clause/T-unit, as well as in clauses per sentence (p. 23-24). At C1 level, Russian and German EFL learners differed in the length of production units and possibly in the proportional indices based on clauses and T-units listed above, as well as in clauses per sentence. Indices of coordination did not appear to vary with L1 at either level. Lu and Ai's study covered a wide range of SC indices and texts, and it suggests that EFL learners' syntax

differs as a function of their L1 even if their CEFR level is the same. However, the fact that an unknown proportion of texts in any L1 group did not belong to the average CEFR level of the group makes these results uncertain.

Two other studies based on other proficiency frameworks than the CEFR have also compared different EFL learners. In an early study, Bardovi-Harlig and Bofman (1989) investigated clauses per T-unit with learners from five L1 backgrounds: Arabic, Chinese, Korean, Malay, and Spanish. Each group included six learners who wrote a composition that required description and possibly some argumentation. The researchers found the clause/T-unit ratio to be similar across L1 groups. Learners' English proficiency was around TOEF score 550 points, which probably corresponds B2 (https://www.etsglobal.org/Tests-Preparation/The-TOEFL-Family-of-Assessments/TOEFL-ITP-Assessment-Series/Scores-Overview). The study is interesting as it covered several very different L1 groups whose proficiency was established with a standardised test. It is obviously limited in terms of the number of SC indices and learners, and by the fact that learners' proficiency was established through an overall proficiency test rather than writing specifically.

Finally, Banerjee *et al.* (2007) examined Chinese (n=159) and Spanish (n=116) IELTS test takers and explored number of dependent clauses per clause and clauses per T-unit by using a writing task requiring expression of opinions with supporting arguments (IELTS writing task 2). SC analyses were based on a sample of 42 texts across both L1 groups and 6 IELTS levels. Findings indicate that neither of the SC indices increased linearly across IELTS levels 3 to 8 in either of the groups. However, clauses per T-unit rose clearly between levels 4 and 5 (roughly A2/B1) among the L1 Spanish while for the Chinese, it started increasing from level 5 onwards and was particularly pronounced between 7 and 8 (B2/C1). The study was based on solid linkage with standardised examination levels but covered only two SC indices and a relatively small number of texts so no statistical analyses were performance on the SC data.

The analysis of previous research indicates, first, that the picture we have about syntactic complexity at different CEFR levels in EFL writing is quite sketchy. Studies that exist have covered somewhat different and often limited sets of indices. Therefore, no clear understanding emerges of the SC features that typically differentiate CEFR levels in EFL learners' writing, apart from the fact that SC usually increases as writing ability improves. Second, studies have covered only one L1 group of EFL learners or a mixture of L1 backgrounds. Hence, little is known how comparable the CEFR levels are across learners who have different first languages, that is, we do not know to what extent previous findings on syntactic complexity have been language-specific rather than general. The very few studies that compare L1 groups are somewhat inconclusive but suggest that learners' L1 might affect their SC. Thirdly, research methods vary considerably, for example, in the number and nature of the writing task: sometimes all participants complete the same writing task(s) under the same circumstances, whereas in other studies learners' texts are less comparable. Some studies have issues with the reliability of placing learners' texts on the CEFR levels. Furthermore, some studies are quite small-scale which makes the (quantitative) analyses less precise.

We will next present our aims and research questions, current study and a description of research methodology: participants, data collection, rating of performances, and analyses. These are followed by the results organised by aspects of SC, and a discussion of the findings with reference to previous research on SC in EFL writing.

AIMS AND RESEARCH QUESTIONS

The present study addresses some of the issues identified in the literature review. It investigates two linguistically different groups of EFL learners in two countries with different cultural, educational and sociolinguistic characteristics (Pakistan with an Indo-Arian language, Sindhi, and Finland with a Finno-Ugric language, Finnish). The learners were in the same age and ability range (from A1 to B1 in EFL writing) and they completed the same writing task under the same conditions. Learners' texts were multiply rated on the CEFR scale and the ratings were analysed to ensure their quality. Thus, the design allows us to investigate syntactic complexity across three CEFR levels in EFL writing, and to find out to what extent the CEFR levels are comparable linguistically across different L1 groups.

We investigate syntactic complexity by using two automated applications developed for analysing English: *L2 Syntactic Complexity Analyzer (L2SCA;* Lu 2010) and *Coh-Metrix* (Graesser *et al.* 2004) which allows us to process the large number of texts involved in the study (about 1,150 texts). We cover almost 30 indices of SC (see Table 2 and 3 in SuppData). There are several reasons for including so many indices. First, complexity is a multidimensional construct, as was described earlier, and so is syntactic complexity. Bulté and Housen (2012) list over 30 SC measures used in research, divisible into at least sentential, clausal and phrasal levels. As our review of CEFR-related SC studies indicates, one of the weaknesses in many studies is the limited range of measures. More generally, too, SLA research on SC has suffered from limited validity as the measured SC construct narrows down because too few indicators are investigated (e.g., Bulté and Housen 2012, 2014, 2018). Secondly, the relationship between different SC indices and L2 proficiency is not clear: the results vary between studies (Lu and Ai 2015). All this speaks for including a wide range of SC indices in research. It should be recognised, however, that many of these measures overlap and tap more than one dimension or level of complexity. Thus, they can be seen as hybrid rather than independent measures of complexity (Bulté and Housen 2012: 10).

The study has two aims: (1) to investigate the linguistic basis of the CEFR levels in EFL writing by examining which syntactic complexity features might *distinguish* different levels, and (2) to examine to what extent SC in EFL might vary across two very different first language groups.

The **research questions (RQ)** were:

1. What syntactic complexity features in argumentative essays written by Sindhi and Finnish EFL learners distinguish between CEFR levels A1, A2 and B1?
2. Which syntactic complexity features differ or remain the same between the Sindhi and Finnish EFL learners when their CEFR writing levels are the same?

METHODOLOGY

Participants

The participants were EFL learners in grades 8-12 from Pakistan and Finland, aged 13–18. There were 868 Sindhi-speaking learners from 31 schools in Pakistan and 287 Finnish-speaking learners from 12 schools in Finland. School selection was based on the researchers' contacts with the schools in the two countries. Different types of schools (city, town, countryside; public, private) were chosen to cover students with a

range of backgrounds. Hence, the Pakistani sample included public (i.e. government) schools (13), as well as private (9) and semi-private (9) schools. With one exception, Sindhi, rather than English, was the medium of instruction in these schools. For Finland, the participating schools were public and the language of instruction was Finnish. The heterogeneity of the educational system in Pakistan (see below), and our desire to cover that variation adequately, were the main reasons for taking a larger sample of students from Pakistan.

The participants represent two very different first languages as well as educational, cultural and sociolinguistic contexts. Typologically, the languages differ, Sindhi being an Indo-Arian (Indo-European) and Finnish being a Finno-Ugric language. English plays an important but different role in both countries. In Pakistan, a former British colony, English is an official language with Urdu and has a very high status. There are also English-medium newspapers and television channels. However, students' proficiency in English is very uneven because of large differences in parents' socio-economic background, the quality and resources available for teaching in schools, and, therefore, access to English both in and out of school (Shamim 2008). According to Rahman (2001: 242), English is a second language for the "affluent, highly educated people and a foreign language for all educated others". In Finland, English has no official status but it is the most popular foreign language that over 90% of secondary level students study. English is very much present in the media (e.g., films are not dubbed) and in young people's free time. Compared to Pakistan, Finnish schools are more homogeneous at least in compulsory education: between-school differences are the smallest among the OECD countries and, thus, the effect of individual schools on outcomes is quite small (e.g. OECD 2016: 226).

Data collection

Data were collected as part of larger studies in which the learners completed several writing tasks in English during their regular lessons. The current study focuses on an argumentative essay in which learners were asked to state their own opinion on a given issue (should mobile phones be allowed in the schools) and give reasons for their opinion (Appendix 1, SuppData). The task elicited, thus, a variety of academic English.

Informed consent was obtained from the students, and the researchers explained task instructions (orally in Sindhi and Urdu in Pakistan; in Finnish in Finland), and supervised task completion. Ample but limited time was given to the participating students to complete the tasks.

Rating procedure

The essays were rated on a six-point scale compiled from several CEFR writing scales (see Huhta et al. 2014). The Finnish scripts had been collected in an earlier project; data collection and rating procedures in Pakistan were modelled on that project. In both countries, the raters were English language experts with master's or doctoral degrees in English. Raters' training sessions comprised an introduction to the scale, rating of sample performances, and discussion of the ratings.

Each Finnish script was judged by 2 raters and each Pakistani script by 4–7 raters; in total, there were 3 Finnish and 14 Pakistani raters. Two of these Finnish raters rated about 30% of the Pakistani scripts to increase the comparability of the assessments. Ratings were analysed with multifaceted Rasch analysis program Facets

(Linacre 2009). The fair average values from Facets were the basis of the placement of the texts on the CEFR levels. Rating quality was controlled with reference to the Infit values (e.g., Engelhard 1994); three misfitting and/or too lenient/severe raters were removed to increase data quality (see Appendix 2 in SuppData for details).

Preparing the corpus

Before automated analyses, corpora are often 'cleaned' to remove issues that can distort the results. No hard and fast guidelines exist but McNamara *et al.* (2014: 155–6) state that when corrections are made they should be carried out systematically. After examining the effect of potentially problematic issues, we corrected minor spelling errors, added missing sentence final punctuation marks, and deleted learners' comments. Extremely short texts (under 10 words), texts written in L1 and texts copied from another student were also removed. We noticed that particularly missing sentence final punctuation affects all SC indices based on sentence length. Apart from spelling errors, other linguistic errors were not corrected.

ANALYSIS

Extraction of syntactic complexity features

Two automated applications were used to extract 28 features to cover the multidimensional SC construct as comprehensively as possible. The first application was *L2 Syntactic Complexity Analyzer (L2SCA;* Lu 2010) and the second was *Coh-Metrix* (Graesser et al. 2004). Tables 2 and 3 in Supplementary Data list all SC indices and define them.

Statistical analyses

We first identified and removed multivariate outliers on Mahalanobis Distance tests in SPSS for groups of SC indices. Descriptive statistics were computed separately for the two language groups (Tables 6-9, Supporting Information). To answer RQ1, a series of MANOVAs were first run for each dimension or combination of dimensions of indices to account for Type I error. These were followed by univariate analyses and pairwise comparisons to determine which indices distinguish the CEFR levels. For RQ2, t-tests were used for comparing the two learner groups.

RESULTS

We first provide an overview of the distribution of the texts across the CEFR levels. Table 2 shows that the number of texts in each category (level / L1) differed; however, even in the smallest category, there were 65 texts.Table 2: Distribution of learners' writings across the CEFR levels in the two countries

Country	A1	A2	B1
Finland	65 (22.7%)	100 (34.8%)	122 (42.5%)
Pakistan	446 (51.4%)	324 (37.3%)	98 (11.3%)

Research Question 1

The results relating the RQ 1 (whether SC indices distinguish the CEFR levels) are presented first, separately for each SC dimension. For convenience, we refer to the two language groups by using the names of the countries they come from (Pakistan and Finland). We display the findings as error-bar charts because they are effective in communicating a large number of comparisons; the detailed descriptive statistics and the numerical results of univariate and pairwise analyses are presented in online Supplementary Data. The error-bar charts also display how the two L1 groups (Sindhi and Finnish) compared but we will give an account of those findings (Research Question 2) only after describing the results related to the CEFR levels.

Length of production units

First, an overall multivariate analysis of the length of production unit indices was conducted; it indicated significant differences across the CEFR levels in both learner groups (Table 4, SuppData). Overall, the mean lengths of the production units distinguished the CEFR levels in both countries and for almost all the three CEFR levels included in the study. Figure 1 shows the error-bar charts for four length measures and display the means and 95% confidence intervals for the two language groups and three CEFR levels in each group (for descriptive statistics and the numerical results of univariate and pairwise analyses, see Table 5 and 9 in SuppData). Particularly sentence and T-unit lengths, and mean standard deviation of sentence length differentiated the CEFR levels; length of clauses did not separate the levels in most cases. The effect sizes (partial eta squares) were high in Finland (e.g., η^2=.248 for sentence length, η^2=.186 for standard deviation of sentence length and η^2=.104 for T-unit length) and with medium effect sizes in Pakistan (highest was η^2=.056 for sentence length). The univariate analyses indicated that separation was clearer between A1 and A2 than between A2 and B1 (i.e., effect sizes were larger for the former).

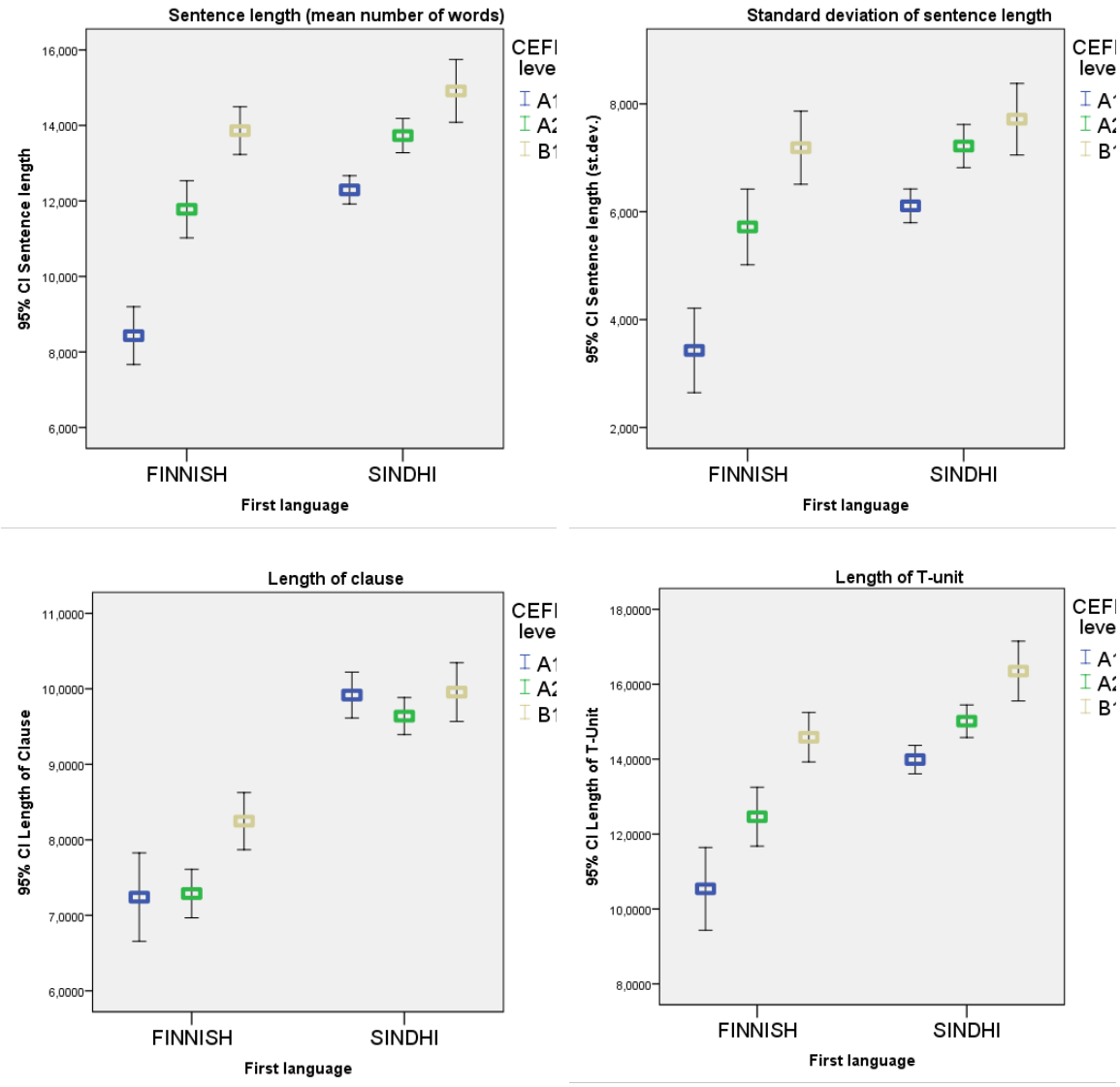


Figure 1. Error-bar charts for differences in the length of production units

Subordination, coordination, and phrasal sophistication

Multivariate analyses indicated significant differences across CEFR levels (Table 4, SuppData). Indices of subordination showed fairly good separation between CEFR levels, particularly for A1 vs B1 but also between adjacent levels (Figure 2; Table 6 and 10, SuppData). Effect sizes ranged only from small to medium, however. The best separators were complex T-units per T-unit and dependent clauses per clause followed by clauses per T-unit; separation was clearer in Finland. Indices of coordination did not separate CEFR levels in either L1 group (Figure 3). Among the indices of phrasal sophistication, verb phrases per T-unit was a significant separator with medium effect size in both countries (Figure 4).

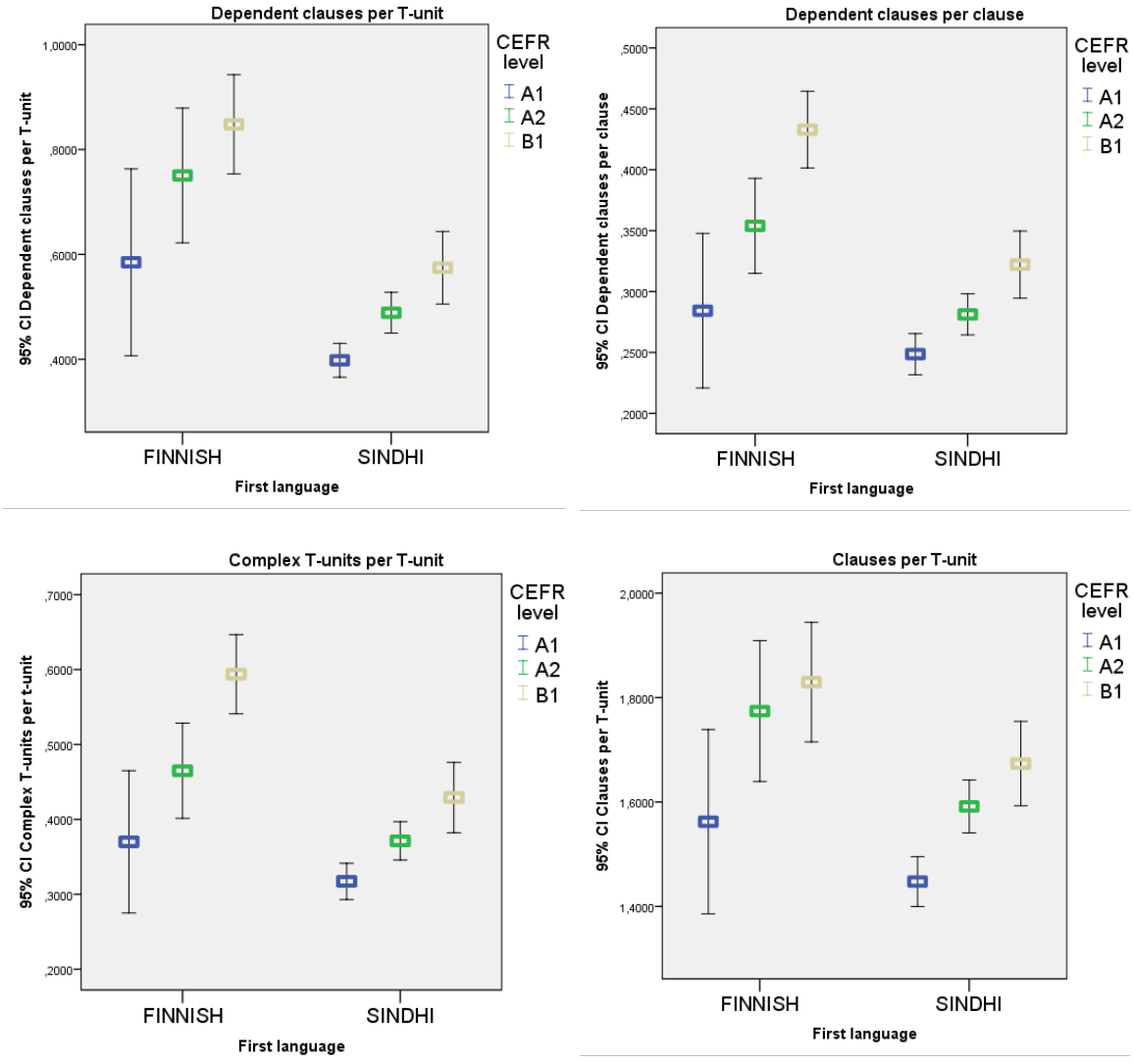


Figure 2. Error-bar charts for differences in subordination

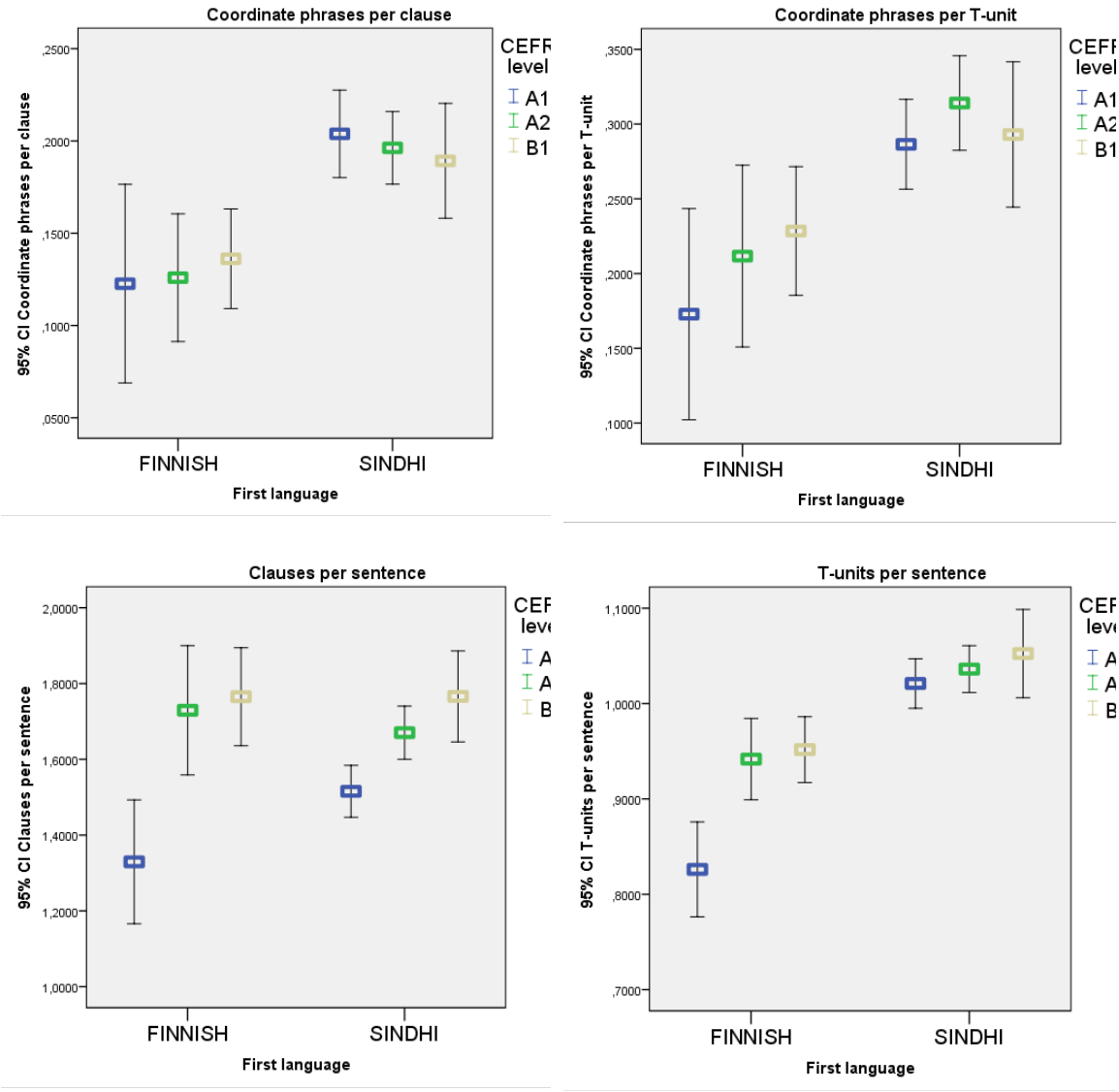


Figure 3. Error-bar charts for differences in coordination

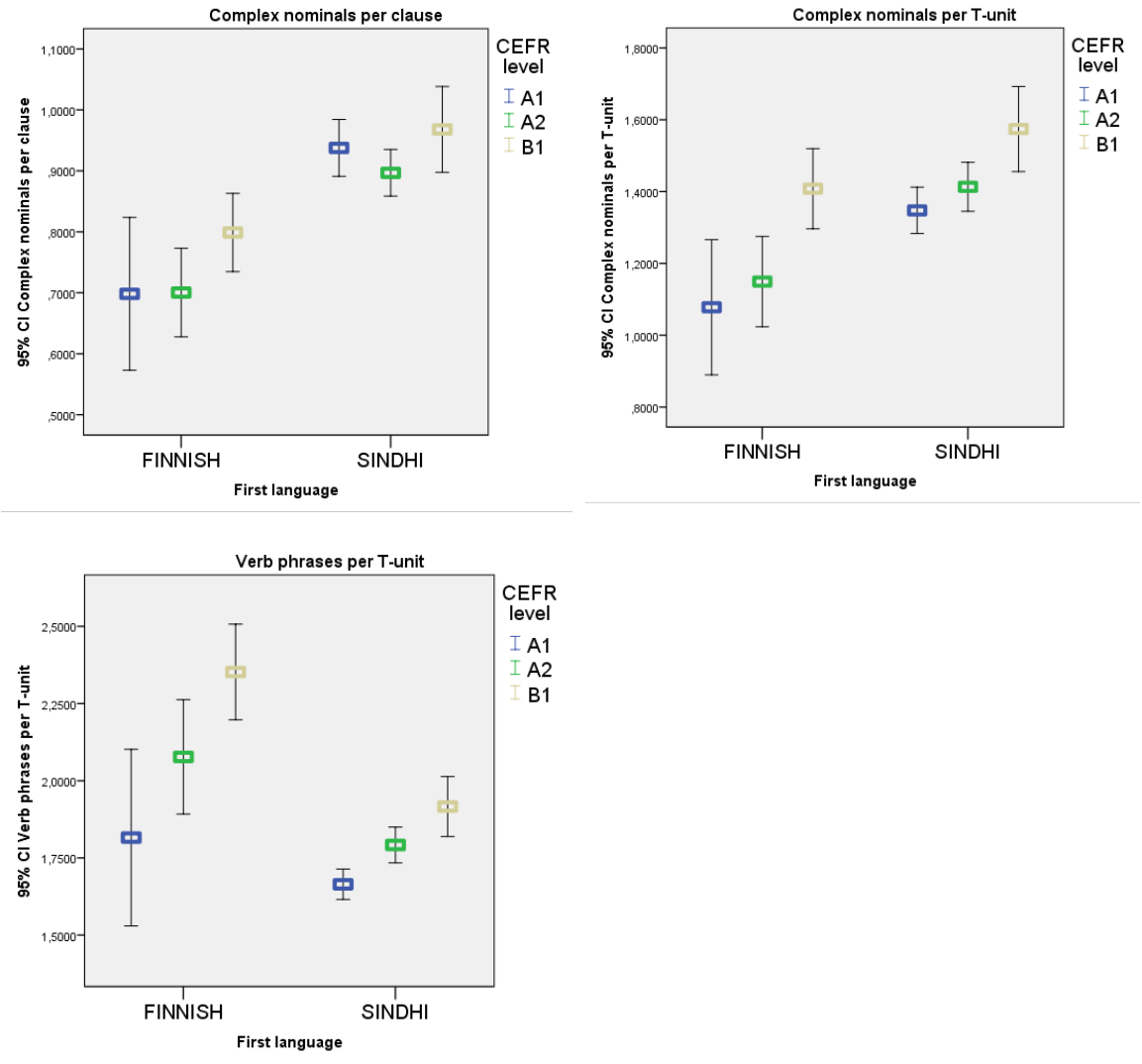


Figure 4. Error-bar charts for differences in phrasal sophistication

Working memory load, referencing expressions and syntactic variability and simplicity

In this group of indices, too, the multivariate analyses demonstrated significant differences across CEFR levels (Table 4, SuppData). Particularly modifiers per noun phrase, left embeddedness, and minimal edit distance (Figure 5) separated CEFR levels, more clearly in Finland (Tables 7 and 11, SuppData). Minimal edit distance, an index of syntactic variety, achieved the highest effect size (η^2=.079) but only among the Finns. Syntactic simplicity z-score, and syntactic structural similarity showed no differences. Modifiers per noun phrases behaved in a different way compared with the other significant SC indices: it exhibited non-linear relationship with the CEFR levels. The values for this index decreased from A1 to A2 (from .215 to .156) but then increased at B1 (from .156 to .217; Table 7, SuppData).

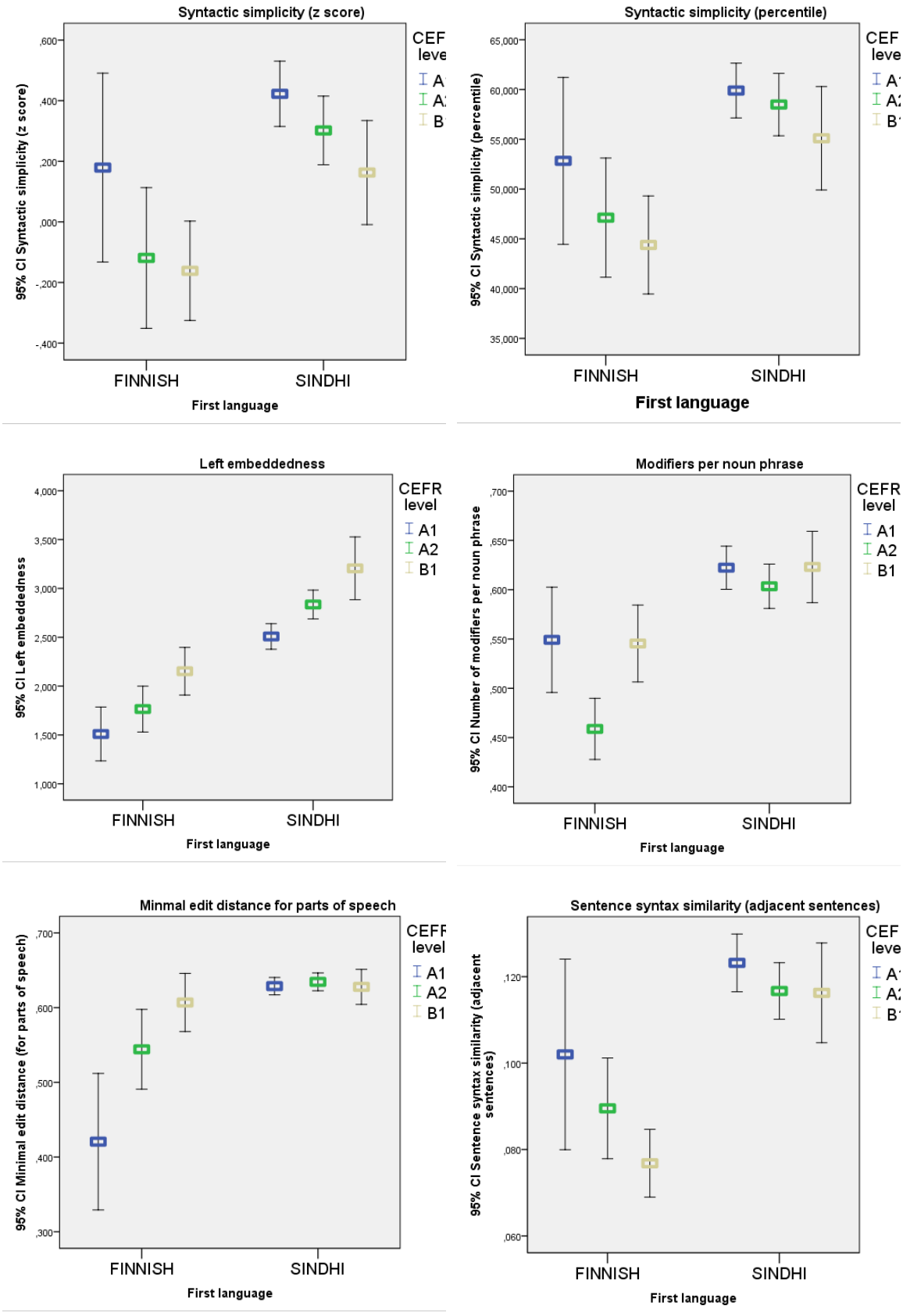


Figure 5. Error-bar charts for differences in working memory load, referencing expressions and syntactic variety and simplicity.

Phrasal density

All phrasal density measures related to SC demonstrated some ability to distinguish CEFR levels in either or both of the countries (Figure 6a/b; Tables 8 and 12, SuppData); multivariate analyses also indicated significant differences across levels (Table 4, SuppData). For the Finns, the best separators were noun phrase and infinitive density with fairly large effect sizes (η^2=.102 and η^2=.081, respectively) followed by gerund and negation densities, and, less so, verb and adverbial phrase densities. In Pakistan, only negation and infinitive density clearly differentiated CEFR levels (with moderate effect sizes; η^2=.043 and η^2=.034, respectively), even if preposition, verb phrase and gerund densities demonstrated some separation. No clear pattern emerged as to whether these indices were better separators in the lower (A1 vs A2) or higher proficiency range (A2 vs B1).

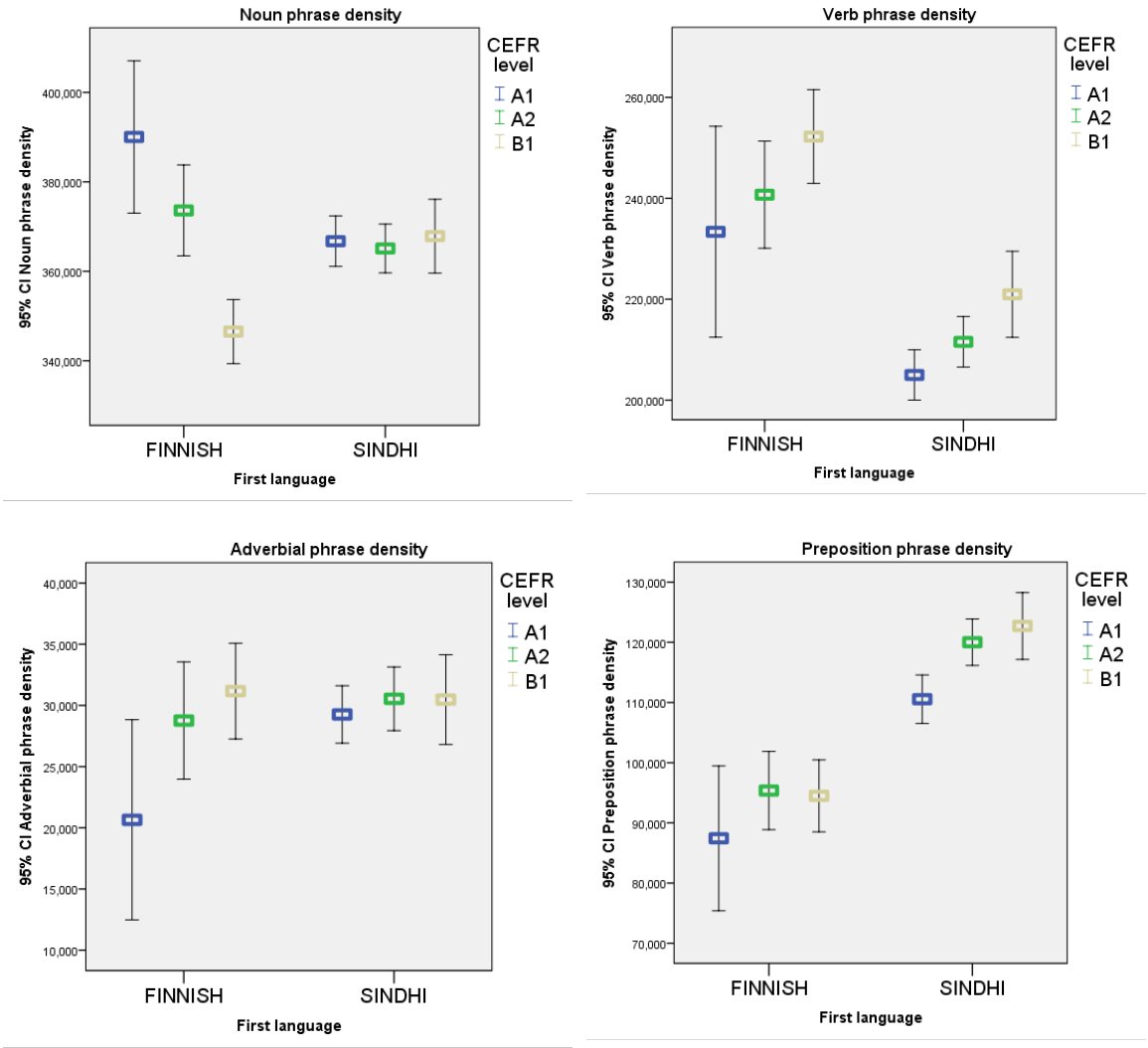


Figure 6a. Error-bar charts for differences in phrasal density

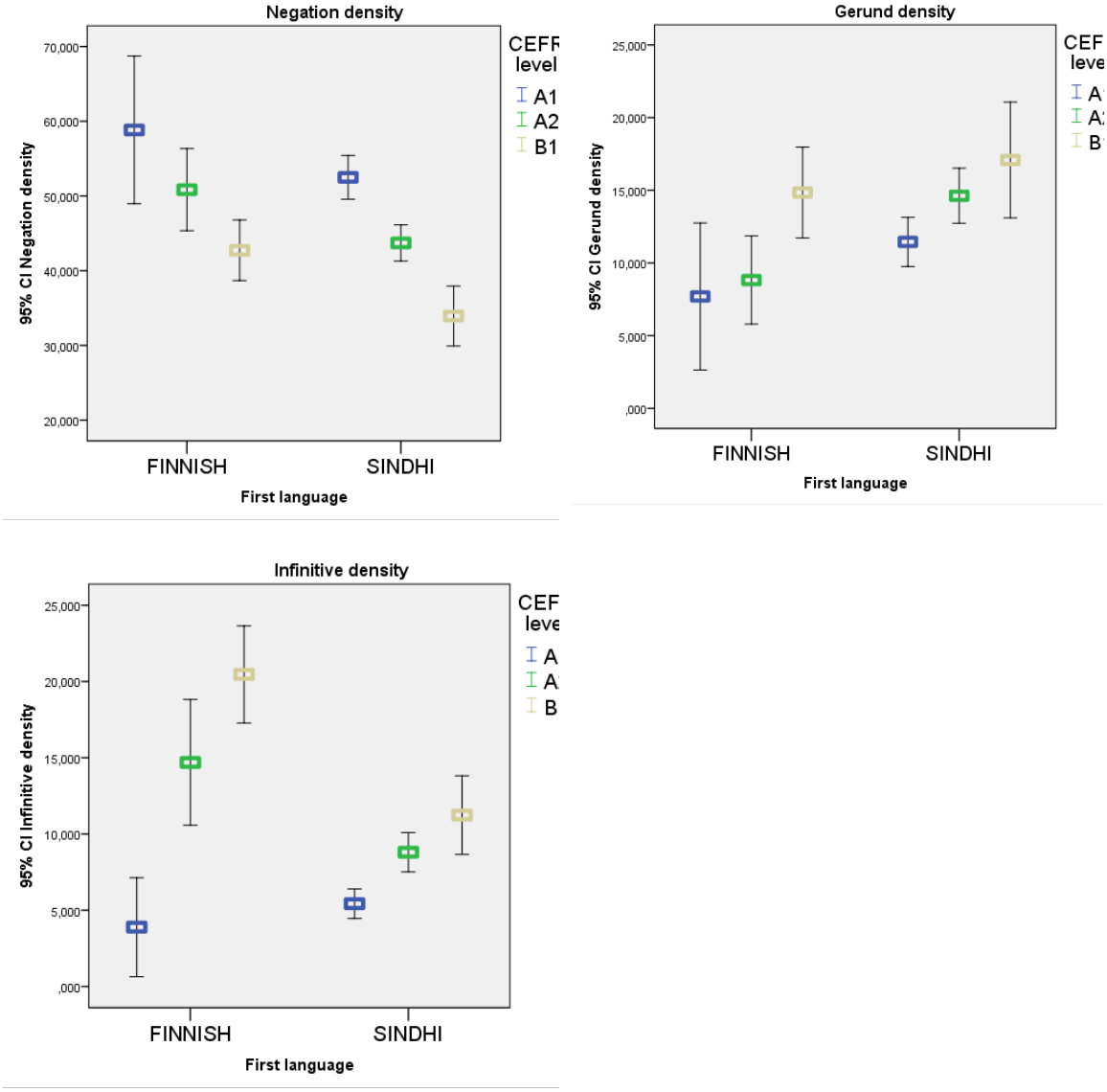


Figure 6b. Error-bar charts for differences in phrasal density

Research Question 2

Our second research question concerned comparability of SC between the Finnish and Sindhi EFL learners whose texts represented the same CEFR levels. Multivariate analyses of all 28 SC indices comparing the L1 groups indicated large overall differences at all three CEFR levels (Table 10, SuppData), which warrants more detailed comparisons.

Figures 1–6 that display differences across CEFR levels also show where similarities and differences between the two L1 groups were found. We summarise these with three tables. Table 3 lists the SC indices that remained the same in both L1 groups whereas Tables 18 and 19 (in Supplementary Data) detail the differences (for exact numerical results, see Tables 14–16, SuppData).

Table 3: Syntactic complexity indices that remained the same across the two language groups.

Finnish A1 vs Sindhi A1	Finnish A2 vs Sindhi A2	Finnish B1 vs Sindhi B1
Verb phrases per T-unit	Clause per T-unit	Clauses per sentence
Syntactic structure similarity	Minimal edit distance	Sentence length (st.dev.)
Syntactic simplicity (z score & percentile)	Noun phrase density	Minimal edit distance
Dependent clauses per T-unit	Clause per sentence	Adverbial phrase density
Complex T-unit per t-unit	Adverbial phrase density (Verb phrases per T-unit)	Gerund density
Coordinate phrases per T-unit		
Coordinate phrases per clause		
Gerund density		
Infinitive density		
Clause per T-unit		
Dependant clause per clause		
(Modifiers per noun phrase)		

Overall, there were more differences in SC between the two L1 groups than there were similarities. Table 3 shows that no index remained the same across all three CEFR levels and only three did so at two levels: clauses per sentence and two density indices (negation and adverbial phrase density), and possibly verb phrases per T-unit. Most SC indices differed significantly between the groups at every CEFR level; the differences were more numerous at A2 and B1 where 22 or 23 of the 28 SC indices separated the L1 groups. Conversely, level A1 was more similar across the two groups than the other levels since the values of as many as 13 of the 28 SC indices were the same. In contrast, only 5 or 6 indices remained the same at A2 and B1.

A closer look at the dimensions/levels of SC reveals that the largest differences occurred in the measures of length of the production unit: the Sindhi-speakers wrote longer sentences, clauses and T-units across all levels (see Figure 1 and Tables 14–16, SuppData). The differences were largest at A1 where the effect sizes (Cohen's d) varied from 1.228 for sentence length to .928 for clause length. Differences were found also at A2 and B1 but with somewhat smaller effect sizes, with clause length being the clearest separator ($d = .986$ at A2; $d = .828$ at B1).

Sindhi-speakers used more coordination (T-units per sentence, coordinate phrases per clause or T-unit), particularly at A1 and A2, whereas Finns used more subordination, especially at A2 and B1 (dependent clauses per clause or T-unit, complex T-units per T-unit). The Coh-Metrix indices of general syntactic similarity and simplicity indicated that Sindhi-speakers' syntax at A2 and B1 was more simple and similar (across sentences) than Finns' syntax.

As to clausal and phrasal sophistication, Sindhi-speakers wrote more complex nominals per clause or per T-unit, and had higher left-embeddedness (more words before main verb) across all CEFR levels. They also used more modifiers per noun

phrase, particularly at A2 and B1. In contrast, Finns used more verb phrases per T-unit but only at B1 (Figure 5; Tables 14–16, SuppData).

The L1 groups also differed at phrasal level (Figure 6a/b). Particularly, preposition phrase density separated at all levels, with the Sindhi-speakers writing denser phrases (d=.511 at A1; d=.871; d=1.030 at B1); their gerund density was also higher at A2. In contrast, in the other large phrasal level separator, verb phrase density, the Finns obtained higher values (d=.413 at A1; d=.554 at A2; d=.784 at B1). The Finn's infinitive and negation phrase densities were also higher, especially at A2 and B1.

DISCUSSION

This study addresses the linguistic basis of the CEFR, which is an important area of investigation given its influence (e.g., Hulstijn 2007; Bartning *et al.* 2010; Wiśniewski 2017). Many SLA studies have examined the relationship between linguistic features and proficiency but few have operationalised proficiency as CEFR levels and, thus, addressed their linguistic characteristics.

We investigated whether syntactic complexity differentiates CEFR levels in EFL learners' writing and whether the results depend on the learners' L1. Thus, the study also sheds light on the linguistic comparability of the CEFR levels. We next discuss our findings with reference to previous research on SC in EFL writing.

Discussion of RQ1: distinguishing CEFR levels

Length of production units

Wolfe-Quintero's (1998) early review indicated that sentence length increases with proficiency and probably differentiates adjacent proficiency levels. In our study, the highest effect size for the differences between the CEFR levels was found for the mean sentence length in the Finnish group (η^2=.235); among the Sindhis, it was somewhat smaller (η^2=.056). Sentence length was the only SC index separating all CEFR levels in both groups (Figure 1). Our finding agrees with Hawkins and Filipovic (2012) who found sentence length to separate all CEFR levels between A2–C2 and with Lahuerta Martínez (2018) who discovered the same for A2 vs B1.

Other length indices also distinguished CEFR levels, particularly among the Finnish learners. These included the standard deviation of sentence length and the mean T-unit length. This is in line with Gyllstad *et al.* (2014) and Verspoor *et al*. (2012) who found T-unit length to distinguish A1 from A2, and A2 from B1. Gyllstad *et al*. (2014) also found mean clause length to distinguish A2 and B1, whereas we found it to be a rather weak separator.

Subordination and coordination

In our study, most subordination indices differentiated between CEFR levels in both countries but more clearly in Finland (Figure 2), and subordination increased with proficiency. Thus, our findings agree with Wolfe-Quintero *et al.* (1998) who argued that, e.g., dependent clauses per clause is an index of language proficiency. They also concur with Kim (2004) who found that subordinate clauses distinguished A2 and B2

and with Lahuerta Martínez (2018) for A2 vs B1. Similarly, Gyllstad et al. (2014) found significant correlations between clauses per T-unit and proficiency.

Coordination indices failed to separate CEFR levels, even though their values increased slightly, particularly between A1 and A2. The exception was number of T-units per sentence, which was a good separator, but only in Finland, and between A1 and A2. Lahuerta Martínez (2018) also found coordination to distinguish A2 from B1.

Phrasal sophistication

Of the indices of phrasal sophistication (Table 2, SuppData), verb phrases per T-unit has been given special attention in previous research but the views about its usefulness differ. Wolfe-Quintero *et al.* (1998: 85, 123) recommended it because it captures both finite and non-finite verb phrases and contributes to the overall measurement of SC. Support comes from Verspoor *et al.* (2012) who discovered that verb phrases per T-unit distinguished certain CEFR levels. In contrast, Lu (2011) found it not to discriminate between the school levels that he used as a proxy for proficiency. In our study, this index turned out to be a good separator in both countries, thus supporting Wolfe-Quintero *et al.* (1998) and Verspoor *et al.* (2012).

Wolfe-Quintero *et al.* (1998) speculated that complex nominals per clause might perform better than complex nominals per T-unit. In our study, however, complex nominals per T-unit was a more consistent separator of proficiency levels in both language groups (Table 6, SuppData). In general, the values for all these indices increased from lower to higher CEFR levels.

Verb and noun phrases

Two Coh-Metrix indices focus on the length of verb and noun phrases. The first is left embeddedness, the number of words before the verb in the main clause of a sentence. It is argued to relate to working memory load: more words before the verb make sentences denser and more ambiguous (Graesser *et al*. 2004). The second is the number of modifiers per noun phrase, considered an index of the complexity of referencing expressions (Weir *et al.* 2013: 504). Green's (2012) study found number of modifiers per noun phrase to rise significantly from B2 to C1. Non-CEFR studies such as Biber *et al.* (2011), Guo *et al.* (2013) and Kyle (2016) have also found proficient writers to produce more complex noun phrases. In our study, the number of modifiers per noun phrase was unique, as it showed non-linear development, first decreasing from A1 to A2 and then increasing from A2 to B1, particularly among the Finns (Figure 5). For left embeddedness, McNamara *et al.* (2010) found it to increase with higher proficiency. Our findings for Sindhi speakers were somewhat similar, as left embeddedness increased and separated A1 from A2 and B1 (but not A2 from B1). On the whole, however, our results for the noun and verb phrase length were quite inconclusive.

Syntactic similarity, variety and simplicity

Coh-Metrix calculates three types of indices that focus on SC from the perspectives of similarity, variety, and simplicity. The only CEFR-related study investigating these indices is Green's (2012) who reported syntactic similarity to decrease as learners' proficiency increased from C1 to C2. In our study, for lower CEFR levels, syntactic similarity of adjacent sentences also decreased, and its counterpart, syntactic variety

(minimal edit distance for parts of speech) increased but only among the Finns, particularly between A1 and A2 (Figure 5).

Phrasal density

Recent research on SC has begun to pay more attention to the phrasal level (Kyle 2016). Consequently, Coh-Metrix incorporates many phrasal density indices (Figure 6a/b; Table 3, SuppData). We found several phrasal indices to separate CEFR levels in one or both language groups. Most indices (infinitive, gerund, preposition, adverbial, and verb phrase densities) increased with proficiency, but negation density decreased. Among the Finns, also noun phrase density decreased, which is at odds with Green's (2012) discovery that it increased between B2 and C1. However, Green's finding concerned higher CEFR levels, which suggests noun phrase development in EFL writing may be nonlinear across the whole CEFR scale or that learners' L1 affects its development.

Both Kim (2004) and we found gerund and infinitive phrases to increase across CEFR levels. In general, in our study, there was a shift from using noun (and negation) phrases towards using various other types of phrases as proficiency increased, particularly among the Finns.

To summarise discussion so far, our study has provided evidence that CEFR levels A1–B1 in EFL writing differ significantly in terms of several dimensions of SC and in two different L1 groups. Length of production units was a particularly robust separator. Also subordination, but not coordination, and phrasal sophistication and density distinguished the levels. Our findings concur with most previous CEFR-related studies but provide a more comprehensive picture across all dimensions of SC.

Discussion of RQ2: similarities and differences between L1 groups

Since our study investigated EFL learners with the same proficiency level but with different L1 backgrounds, the results shed light on the linguistic generalisability of the CEFR levels.

In general, only some SC indices turned out to be similar across both L1 groups (Table 3). At level A1, 12 of the 28 indices were similar, but as proficiency grew, linguistic differences also grew, and at A2 and at B1 only 5–6 indices remained the same. This pattern suggests that level A1 is more comparable in terms of SC in EFL writing across L1 groups than the subsequent CEFR levels. Level A1 seems to differ from the two higher levels also when we focus on SC indices that did not change with learners' L1: almost all the similarities were unique to A1. The only exceptions were clauses per T-unit and verb phrases per T-unit (shared with A2), and gerund density (shared with B1).

Our findings suggest that, in the A1–B1 range at least, the CEFR scale is most generalizable across languages at A1. The most similar indices (the most overlapping error-bars in Figures 1–6) concerned sentence level similarity and simplicity, subordination, and certain phrasal indices. These aspects and indices may, thus, be more generalizable across languages at the lowest CEFR level than other features of SC.

Beyond A1, however, most SC measures differed significantly across the L1 groups (Tables 18–19, SuppData). The most notable trend concerned the length of the production units: Sindhi-speakers wrote clearly longer sentences, clauses and T-units than their equally proficient Finnish peers. Length differences were most pronounced at A1 but continued at A2–B1. Sindhi-speakers used more coordination whereas Finns

used more subordination in their EFL writing; in general, Sindhi-speakers sentences were simpler, which may be linked with their preference for coordination. They also used more similar sentences across their text. Typical of Sindhi-speakers writing was complexity of noun phrases and a greater number of nominals per clause or T-unit, as well as density of preposition phrases and left-embeddedness (words before verb). These phrasal level characteristics probably explain why Sindhi-speakers' clauses and sentences were longer.

Some characteristics of English spoken in Pakistan may explain why the Pakistani students wrote longer phrases, clauses and sentences. An example is their tendency to use the (longer) perfective aspect instead of the simple past (e.g. 'I *have seen* him yesterday' instead of 'I *saw* him yesterday'; Khan 2012). A possible reason for the finding concerning left-embeddedness may be that because Sindhi is a Subject-Object-Verb language (SOV) its L1 speakers may place more of the sentence elements before the verb when using a foreign language compared to SVO languages such as Finnish (see also Lashari and Soomro 2013). However, unknown differences in teaching methods and materials may also contribute to these differences.

The main conclusion from the above discussion of RQ2 is that the three lowest CEFR levels, particularly A2 and B1 are not comparable with respect to syntactic complexity in EFL writing between L1 speakers of Sindhi and Finnish. This suggests that some, perhaps all, CEFR levels are not equivalent linguistically and, therefore, the development of descriptors, teaching materials and assessments for syntactic complexity needs to consider not only the target language but also learners' L1.

Furthermore, research on the linguistic basis of the CEFR levels may contribute to the investigation of the relationship between different writing and speaking scales. Table 17 in Supplementary Data illustrates how the CEFR and IELTS scales align themselves with respect to two SC indices that were included in Banerjee *et al.* (2007) and in our study. Obviously, proper comparison would require a more extensive comparison of linguistic indices but Table 17 exemplifies the principle.

Overall, the study exemplifies research called for by investigators advocating studies that combine language testing and SLA approaches (Bachman and Cohen 1998), particularly with reference to the CEFR (e.g., Hulstijn *et al.* 2010). We applied procedures developed in language testing to ensure reliable placement of writing samples to proficiency levels to address questions of interest to SLA research and the CEFR. In turn, these findings can help language assessment professionals develop more nuanced understandings of proficiency levels, which is essential for the designing assessments and interpreting their results with respect to specific levels and learners representing particular L1 backgrounds.

CONCLUSION

This study addressed the linguistic basis of the CEFR by focusing on syntactic complexity in Sindhi and Finnish EFL learners' writing. We investigated differences between CEFR levels and compared the two L1 groups to examine whether the findings depend on learners' L1. Most SC indices were found to differentiate CEFR levels in both groups. However, the results varied depending on learners' L1, which suggests that the CEFR levels A1–B1 are not comparable with respect to SC.

The study was limited to one writing task and one pair of L1s, and covered only levels A1–B1. Studies using several tasks, first languages and CEFR levels are needed to obtain a fuller picture of the relationship between SC and CEFR levels. Furthermore,

since writing development is typically heavily influenced by teaching and teaching materials, studies investigating school-aged learners should examine their education in enough detail to establish how syntax is taught at school. This can also help disentangle differences in SC due to learners' L1 from those arising from teaching. Finally, as was discussed earlier, indices of syntactic complexity represent absolute, objective complexity whereas scales such as the CEFR may have more to do with degrees of difficulty of processing and learning (i.e., relative complexity). How these two types of complexity relate is a theoretical challenge but empirical research like the current study might also contribute to the conceptual discussions about complexity.

REFERENCES

Alderson, C. 2007. 'The CEFR and the need for more research,' *Modern Language Journal* 91: 659–663.

Alexopoulou, T., Michel, M., Murakami, A., Meurers, D. 2017. 'Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques,' *Language Learning* 67: 180–208.

Bachman, L., Cohen, A. 1998: 'Language testing–SLA interfaces: an update'. In Bachman, L., Cohen, A. (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research*. CUP.

Banerjee, J., Franceschina, F., Smith, A. 2007. 'Documenting features of written language production typical of different IELTS band score levels,' *IELTS Research Reports, Vol. 7*. IELTS Australia and British Council.

Bartning, I., Martin, M., Vedder, I. 2010. *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research.* Eurosla.

Biber, D., Gray, B., Poonpon, K. 2011. 'Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?,' *TESOL Quarterly* 45: 5–35.

Bulté, B., Housen, A. 2012. 'Defining and operationalising L2 complexity,' In Housen, A., Kuiken, F., Vedder, I. (eds.). *Dimensions of L2 Performance and Proficiency. Investigating Complexity, Accuracy and Fluency in SLA*. Benjamins, pp. 21–46.

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

Bulté, B., Housen, A. 2018. 'Syntactic complexity in L2 writing: Individual pathways and group trends,' *International Journal of Applied Linguistics* 28: 147–164.

Carlsen, C. 2012. 'Proficiency level – a fuzzy variable in computer learner corpora,' *Applied Linguistics* 33: 161–183.

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* CUP.

Dahl, Ö. 2004. *The growth and maintenance of linguistic complexity*. John Benjamins.

Engelhard, G. 1994. 'Examining rater errors in the assessment of written composition with a many-faceted Rasch model,' *Journal of Educational Measurement* 31: 93–112.

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

Green, A. 2012. *Language Functions revisited: Theoretical and Empirical Bases for Language Construct Definition across the Ability Range*. CUP.

Guo, L., Crossley, S., McNamara, D. 2013. 'Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study,' *Assessing Writing* 18: 218–238.

Gyllstad, H., Granfeldt, J., Bernardini, P., Källkvist, M. 2014. 'Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian,' *EUROSLA Yearbook* 14: 1–30.

Hawkins, J., Filipović, L. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework.* CUP.

Housen, A., Simoens, H. 2016 'Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition,' *Studies in Second Language Acquisition* 38: 163–175

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., Hirvelä, T. 2014. Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing 31* 307–328.

Hulstijn, J. 2007. 'The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency,' *Modern Language Journal* 91: 663–667.

Hulstijn, J., Alderson, C, Schoonen, R. 2010. 'Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them,' In Bartning, I., Martin, M., Vedder, I. (eds.). *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*. Eurosla, pp. 11–20.

Khan, H. 2012. 'The evolution of Pakistani English (PakE) as a legitimate variety of English,' *International Journal of Applied Linguistics & English Literature* 1: 90–99.

Kim, S. 2004. *A Study of Development in Syntactic Complexity by Chinese Learners of English and its Implications on the CEF Scales.* MA dissertation. Lancaster University.

Kyle, K. 2016. *Measuring Syntactic Development in L2 Writing: Fine-grained Indices of Syntactic Complexity and Usage-based Indices of Syntactic Sophistication.* PhD Dissertation. Georgia State University.

Lahuerta Martínez, A. 2018. 'Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels,' *Assessing Writing* 35: 1–11

Lashari, M., Soomro, A, 2013. 'Subject-verb agreement in Sindhi and English: A comparative study.' *Language in India* 13: 473-495.

Linacre, M. 2009. *A User's Guide to FACETS v 3.66.0*. Winsteps.

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

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

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

McNamara, D., Crossley, S., McCarthy, P. 2010. 'Linguistic features of writing quality,' *Written Communication* 27: 57–86.

McNamara, D., Graesser, A., McCarthy, P., Cai, Z. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. CUP.

Miestamo, M. 2008. 'Grammatical complexity in a cross-linguistic perspective,' In Miestamo, M., Sinnemäki, K., Karlsson, F. (Eds.), Language Complexity: Typology, Contact, Change. John Benjamins, pp. 23–41.

Kusters, W. 2008. 'Complexity in linguistic theory, language learning and language change,'. In Miestamo, M., Sinnemäki, K., Karlsson, F. (eds.). *Language Complexity: Typology, Contact, Change*. John Benjamins, pp. 3–22.

Milton, J. 2013. 'Measuring the contribution of vocabulary knowledge to proficiency in the four skills,' In Bardel, C., Lindqvist, C., Laufer, B. (eds.). *L2 Vocabulary Acquisition, Knowledge and Use*. Eurosla, pp. 57–78.

Norris, J., Ortega, L. 2009. 'Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA,' *Applied Linguistics* 30: 555–578.

OECD 2016. *PISA 2015 Results (Volume I): Excellence and Equity in Education.* OECD Publishing.

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

Rahman, T. 2001. 'English-teaching institutions in Pakistan,' *Journal of Multilingual and Multicultural Development* 22: 242–261.

Shamim, F. 2008. 'Trends, issues and challenges in English language education in Pakistan,' *Asia Pacific Journal of Education* 28: 235–249.

Treffers-Daller, J., Parslow, P., Williams, S. 2016. 'Back to basics: How measures of lexical diversity can help discriminate between CEFR levels'. *Applied Linguistics*, amw009, https://doi-org.ezproxy.jyu.fi/10.1093/applin/amw009.

Verspoor, M., Schmid, M. S., Xu, X. 2012. 'A dynamic usage-based perspective on L2 writing,' *Journal of Second Language Writing* 21: 239–263.

Weir, C. 2013. 'The measurement of writing ability 1913 – 2012'. In Weir, C., Vidakovic, I., Galaczi, E., 2013 *Measured Constructs: A History of the Constructs Underlying Cambridge English Language ESOL Examinations 1913-2012*. UCLES/CUP.

Wisniewski, K. 2017, 'Empirical Learner Language and the Levels of the *Common European Framework of Reference'*. Language Learning, 67: 232–253.

Wolfe-Quintero, K., Inagaki, K., & Kim, H.-Y. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. University of Hawaii Press.

SUPPLEMENTARY DATA

TABLES

Table 1 Supplementary Data: References to complexity in the CEFR scales

CEFR	References to complexity used in scales of CEFR
C2	clear, complex, logical structure (p.61), a wide range of language (p.110), complex language (p. 28, 114), complex reports, articles or essays (p.62)
C1	clear, well-structured texts of complex subjects (p.61), appropriate formulation (p.110), a broad range of language(p.110), clear, well-structured expositions of complex subjects (p.62)
B2	clear, detailed texts on a variety of subjects (p.61), longer complex stretches of speech (p.129), some complex sentence forms (p.110)
B1	straightforward ….texts (p.61), a repertoire of frequently used patterns (p.114), expressing more complex thoughts (p.112), Simple face-to-face conversation (p. 86,124), can write short, simple essays, brief reports (p.62)
A2	simple phrases and sentences (p.61), Simple face-to-face conversation (p.86,124), basic language(p.110, 122), simple structures (p.29, 114), simple sentences, (p.125), basic sentence patterns (p. 29, 110)
A1	Simple isolated phrases and sentences, (p.61), basic range of simple expressions" (p.110), simple sentence patterns(p.114)

Table 2 Supplementary Data: Syntactic complexity indices calculated by L2 Syntactic Complexity Analyzer

Dimensions	Indices (Lu and Ai 2015: 18)
Length of production units (in words)	Sentence length, T-unit length, Clause length (Standard) deviation of the mean sentence length was calculated with Coh-Metrix)
Amount of subordination	Clauses per T-unit, Dependent clauses per clause, Complex T-units per T-unit, Dependent clauses per T-unit
Amount of coordination	T-unit per sentence, Coordinate phrases per clause, Coordinate phrases per T-unit
Degree of phrasal sophistication	Complex nominals per clause, Complex nominals per T-unit, Verb phrases per T-unit
Sentence complexity	Clauses per sentence

T-unit = "is defined the unit generated when text is divided into the smallest possible independent segments, without leaving sentence fragments behind. Each T-unit consists of a main clause and all the subordinate clauses that belong to it." (Banerjee et al. 2007: 41)

Table 3 Supplementary Data: Syntactic complexity indices calculated by Coh-Metrix

Dimensions	Indices
Working memory load	Left embeddedness (number of words before the verb in the main clause)
Referencing expressions	Modifiers per noun phrase
Phrasal density	Noun phrase density, Preposition phrase density, Verb phrase density, Adverbial phrase density, Infinitive density, Gerund density & Negation density (incidences per 1000 words)
Simplicity	Syntactic simplicity (percentile and Z-score; Degree to which sentences in the text contain fewer vs more words and use simple vs complex syntactic structures)
Similarity	Syntactic structure similarity (across adjacent sentences)
Variety	Minimal edit distance (for parts of speech)

Table 4 Supplementary Data: MANOVA analyses of the differences between CEFR levels A1, A2 and B1 in Finnish (Finnish) and Sindhi (Sindhi) across dimensions of syntactic complexity

Dimension of SC / Group of SC indices		Wilks's Lambda	*F*	*p*	η^2
Length of production units	Finnish	.649	16.183	<.001	.195
	Sindhi	.930	7.672	<.001	.035
Subordination, coordination, and particular structures	Finnish	.704	4.484	<.001	.161
	Sindhi	.940	2.337	<.001	.031
Working memory load, referencing expressions, syntactic variability and simplicity	Finnish	.771	5.366	<.001	.122
	Sindhi	.967	1.949	.018	.016
Phrasal density	Finnish	.797	4.633	<.001	.107
	Sindhi	.898	6.632	<.001	.052

Table 5 Supplementary Data: Pairwise and overall differences of syntactic complexity features across CEFR levels in Finnish and Sindhi: Length of production units

		Pairwise			Overall		
Syntactic complexity indices	Language	A1 vs A2 *p*-value	A2 vs B1 *p*-value	A1 vs B1 *p*-value	*F*	*p*	*η2*
Mean length of sentence	Finnish	0.000	0.000	0.000	44.55	0.000	0.248
	Sindhi	0.000	0.020	0.000	25.03	0.000	0.056
Sentence length (S.D)	Finnish	0.000	0.000	0.000	30.97	0.000	0.186
	Sindhi	0.001	0.589	0.001	10.65	0.000	0.025
Mean length of T-unit	Finnish	0.143	0.001	0.000	15.68	0.000	0.104
	Sindhi	0.000	0.017	0.000	17.83	0.000	0.041
Mean length of clause	Finnish	0.724	0.064	0.008	5.31	0.005	0.038
	Sindhi	0.070	0.637	0.539	2.33	0.098	0.006

Table 6 Supplementary Data: Pairwise and overall differences of syntactic complexity features across CEFR levels in Finnish and Sindhi: Subordination, coordination, sentence complexity and phrasal sophistication indices

		Pairwise			Overall		
Syntactic complexity indices	Language	A1 vs A2 *p-value*	A2 vs B1 *p-value*	A1 vs B1 *p-value*	*F*	*p*	*η2*
Clauses per sentence	Finnish	0.001	0.449	0.000	12.21	0.000	0.084
	Sindhi	0.031	0.563	0.012	5.83	0.003	0.014
Verb phrases per T-unit	Finnish	0.015	0.001	0.000	18.87	0.000	0.124
	Sindhi	0.001	0.019	0.000	16.10	0.000	0.038
Clauses per T-unit	Finnish	0.072	0.365	0.001	6.54	0.002	0.047
	Sindhi	0.001	0.275	0.000	12.20	0.000	0.029
Dependant clauses per clause	Finnish	0.082	0.004	0.000	14.27	0.000	0.097
	Sindhi	0.027	0.037	0.000	8.96	0.000	0.021
Dependant clauses per T-unit	Finnish	0.092	0.066	0.000	9.15	0.000	0.064
	Sindhi	0.009	0.082	0.000	10.62	0.000	0.025
T-units per sentence	Finnish	0.008	0.994	0.003	6.27	0.002	0.045
	Sindhi	0.999	1.000	0.997	0.02	0.978	0.000
Complex T-units per T-unit	Finnish	0.159	0.002	0.000	13.81	0.000	0.094
	Sindhi	0.039	0.135	0.001	7.90	0.000	0.019
Coordinate phrases per T-unit	Finnish	0.185	1.000	0.092	2.01	0.136	0.015
	Sindhi	0.219	1.000	0.627	1.73	0.179	0.004
Coordinate phrases per clause	Finnish	0.399	1.000	0.384	1.25	0.288	0.009
	Sindhi	0.974	0.654	0.783	0.22	0.803	0.001
Complex nominals per T-unit	Finnish	0.129	0.064	0.000	8.62	0.000	0.061
	Sindhi	0.099	0.046	0.000	8.13	0.000	0.019
Complex nominals per clause	Finnish	0.903	0.112	0.141	2.96	0.054	0.022
	Sindhi	0.749	0.275	0.610	0.84	0.430	0.002

Table 7 Supplementary Data: Pairwise and overall differences of syntactic complexity features across CEFR levels in Finnish and Sindhi: Syntactic simplicity, similarity, variety, working memory load and referencing expression indices

		Pairwise			Overall		
Syntactic complexity indices	Language	A1 vs A2 *p-value*	A2 vs B1 *p-value*	A1 vs B1 *p-value*	*F*	*p*	*η2*
Syntactic simplicity (z-score)	Finnish	0.297	0.999	0.255	1.72	0.181	0.012
	Sindhi	0.959	0.175	0.108	1.68	0.187	0.004
Syntactic simplicity (percentile)	Finnish	0.604	0.882	0.348	1.11	0.332	0.008
	Sindhi	0.999	0.372	0.391	1.17	0.310	0.003
Left embeddedness (words before main verb)	Finnish	0.887	0.480	0.198	1.84	0.161	0.013
	Sindhi	0.035	0.159	0.001	7.79	0.000	0.019
Modifiers per noun phrase	Finnish	0.026	0.002	0.993	6.00	0.003	0.042
	Sindhi	0.718	0.898	1.000	0.49	0.611	0.001
Minimal edit distance (parts of speech)	Finnish	0.036	0.088	0.000	11.88	0.000	0.079
	Sindhi	0.646	0.887	0.990	0.38	0.682	0.001
Sentence syntax similarity (adjacent sentences)	Finnish	0.562	0.268	0.117	3.23	0.041	0.023
	Sindhi	0.807	0.684	0.395	0.74	0.479	0.002

Table 8 Supplementary Data: Pairwise and overall differences of Syntactic Complexity features across CEFR levels in Finnish and Sindhi: Phrasal density indices

		Pairwise			Overall		
Syntactic complexity indices	Language	A1 vs A2 *p-value*	A2 vs B1 *p-value*	A1 vs B1 *p-value*	*F*	*p*	*η2*
Noun phrase density	Finnish	0.437	0.000	0.000	15.65	0.000	0.102
	Sindhi	0.799	0.852	0.999	0.23	0.794	0.001
Verb phrase density	Finnish	0.591	0.172	0.082	3.58	0.029	0.025
	Sindhi	0.169	0.094	0.003	5.28	0.005	0.012
Adverbial phrase density	Finnish	0.415	0.704	0.162	2.40	0.092	0.017
	Sindhi	0.727	0.997	0.849	0.33	0.718	0.001
Preposition phrase density	Finnish	0.887	0.987	0.928	0.14	0.868	0.001
	Sindhi	0.004	0.762	0.003	6.89	0.001	0.016
Negation density	Finnish	0.355	0.067	0.012	6.31	0.002	0.044
	Sindhi	0.001	0.000	0.000	19.14	0.000	0.043
Gerund density	Finnish	0.909	0.031	0.014	5.35	0.005	0.037
	Sindhi	0.011	0.548	0.005	7.15	0.001	0.017
Infinitive density	Finnish	0.045	0.022	0.000	12.18	0.000	0.081
	Sindhi	0.000	0.191	0.000	14.95	0.000	0.034

Table 9 Supplementary Data: Characteristics of CEFR levels in Finnish and Sindhi: Means, standard deviations, and confidence intervals for length of production units

Syntactic complexity indices		A1			A2			B1		
	Language	M	S.D	[95% CI]	M	S.D	[95% CI]	M	S.D	[95% CI]
Mean length of Sentence	Finnish	8.61	3.54	[7.72, 9.50]	11.58	3.61	[10.84, 12.31]	13.80	3.68	[13.16, 14.45]
	Sindhi	12.10	3.75	[11.74, 12.47]	13.56	3.99	[13.13, 13.98]	14.78	3.93	[14.01, 15.56]
Sentence length S.D	Finnish	3.29	2.99	[2.50, 4.08]	5.47	2.78	[4.81, 6.12]	7.18	3.62	[6.60, 7.77]
	Sindhi	6.19	3.45	[5.87, 6.52]	7.15	3.55	[6.77, 7.53]	7.60	3.03	[6.91, 8.29]
Mean length of T-Unit	Finnish	11.10	5.32	[10.05, 12.14]	12.45	4.02	[11.59, 13.31]	14.62	3.71	[13.84, 15.39]
	Sindhi	13.84	3.85	[13.47, 14.20]	14.96	3.93	[14.53, 15.39]	16.21	3.81	[15.43, 16.99]
Mean length of clause	Finnish	7.26	2.37	[6.70, 7.82]	7.60	2.27	[7.15, 8.07]	8.33	2.22	[7.92, 8.75]
	Sindhi	10.25	3.81	[9.95, 10.55]	9.74	2.39	[9.39, 10.09]	9.96	1.94	[9.33, 10.59]

Table 10 Supplementary Data: Characteristics of CEFR levels in Finnish and Sindhi: Means, standard deviations, and confidence intervals for subordination coordination, and sentence complexity

Syntactic complexity indices		A1			A2			B1		
		M	S.D	[95% CI]	M	S.D	[95% CI]	M	S.D	[95% CI]
Clauses per sentence	Finnish	1.27	0.55	[1.11, 1.42]	1.63	0.63	[1.51, 1.75]	1.74	0.59	[1.63, 1.85]
	Sindhi	1.53	0.64	[1.47, 1.59]	1.65	0.60	[1.58, 1.72]	1.73	0.54	[1.61, 1.85]
Verb phrases per T-unit	Finnish	1.60	0.81	[1.42, 1.79]	1.95	0.64	[1.80, 2.10]	2.31	0.76	[2.17, 2.44]
	Sindhi	1.63	0.45	[1.59, 1.68]	1.76	0.47	[1.71, 1.81]	1.91	0.48	[1.81, 2.00]
Clauses per T-unit	Finnish	1.49	0.55	[1.34, 1.64]	1.71	0.57	[1.59, 1.83]	1.83	0.61	[1.72, 1.94]
	Sindhi	1.46	0.43	[1.42, 1.50]	1.58	0.42	[1.53, 1.63]	1.66	0.39	[1.58, 1.74]
Dependant clauses per clause	Finnish	0.27	0.24	[0.22, 0.32]	0.35	0.19	[0.31, 0.39]	0.43	0.16	[0.39, 0.46]
	Sindhi	0.25	0.17	[0.23, 0.26]	0.28	0.15	[0.26, 0.30]	0.32	0.14	[0.29, 0.35]
Dependant clauses per T-unit	Finnish	0.49	0.55	[0.36, 0.62]	0.68	0.51	[0.57, 0.78]	0.84	0.52	[0.75, 0.94]
	Sindhi	0.40	0.33	[0.37, 0.44]	0.48	0.34	[0.44, 0.52]	0.56	0.33	[0.50, 0.63]
T-units per sentence	Finnish	0.85	0.18	[0.80, 0.90]	0.96	0.23	[0.91, 1.00]	0.96	0.20	[0.92, 1.00]
	Sindhi	1.04	0.27	[1.01, 1.06]	1.04	0.23	[1.01, 1.07]	1.04	0.21	[0.99, 1.09]
Complex T-units per T-unit	Finnish	0.35	0.36	[0.27, 0.43]	0.45	0.31	[0.39, 0.51]	0.60	0.29	[0.54, 0.65]
	Sindhi	0.32	0.26	[0.30, 0.35]	0.37	0.23	[0.34, 0.39]	0.42	0.23	[0.37, 0.47]
Coordinate phrases per T-unit	Finnish	0.14	0.23	[0.07, 0.21]	0.22	0.31	[0.17, 0.27]	0.22	0.23	[0.17, 0.27]

	Sindhi	0.27	0.29	[0.24, 0.29]	0.30	0.28	[0.27, 0.34]	0.30	0.26	[0.24, 0.36]
Coordinate phrases per clause	Finnish	0.09	0.15	[0.05, 0.13]	0.13	0.18	[0.10, 0.16]	0.13	0.14	[0.10, 0.16]
	Sindhi	0.20	0.24	[0.18, 0.22]	0.20	0.19	[0.18, 0.23]	0.18	0.15	[0.14, 0.23]
Complex nominals per T-unit	Finnish	0.97	0.62	[0.79, 1.14]	1.19	0.75	[1.06, 1.33]	1.41	0.63	[1.28, 1.53]
	Sindhi	1.30	0.64	[1.24, 1.36]	1.40	0.61	1.33, 1.47]	1.58	0.59	[1.45, 1.70]
Complex nominals per clause	Finnish	0.67	0.46	[0.57, 0.77]	0.70	0.37	[0.62, 0.78]	0.80	0.35	[0.73, 0.87]
	Sindhi	0.93	0.49	[0.89, 0.98]	0.91	0.38	[0.86, 0.96]	0.97	0.35	[0.89, 1.06]

Table 11 Supplementary Data: Characteristics of CEFR levels in Finnish and Sindhi: Means, standard deviations, and confidence intervals for syntactic simplicity, similarity, variety working memory load and referencing expression indices

Syntactic complexity indices		A1			A2			B1		
	Language	M	S.D	[95% CI]	M	SD	[95% CI]	M	SD	[95% CI]
Syntactic simplicity (z score)	Finnish	0.10	1.21	[-0.15, 0.37]	-0.17	1.08	[-0.38, 0.04]	-0.16	0.92	[-0.36, 0.02]
	Sindhi	0.36	1.06	[0.26, 0.46]	0.34	0.99	[0.23, 0.45]	0.15	0.87	[-0.05, 0.36]
Syntactic simplicity (percentile)	Finnish	51.37	33.33	[44.07, 58.67]	46.43	29.44	[40.55, 52.31]	44.57	27.12	[39.26, 49.88]
	Sindhi	59.30	29.19	[56.57, 62.03]	59.54	27.95	[56.40, 62.68]	54.68	26.12	[48.96, 60.39]
Left embeddedness (words before main verb)	Finnish	1.77	1.60	[1.41, 2.13]	1.92	1.45	[1.63, 2.21]	2.18	1.35	[1.92, 2.44]
	Sindhi	2.62	1.53	[2.47, 2.76]	2.90	1.48	[2.73, 3.07]	3.24	1.61	[2.94, 3.55]
Modifiers per noun phrase	Finnish	0.54	0.21	[0.49, 0.59]	0.46	0.16	[0.42, 0.50]	0.55	0.21	[0.51, 0.58]
	Sindhi	0.62	0.23	[0.60, 0.64]	0.61	0.21	[0.58, 0.63]	0.62	0.18	[0.58, 0.67]
Minimal edit distance (parts of speech)	Finnish	0.42	0.37	[0.35, 0.49]	0.55	0.26	[0.50, 0.61]	0.62	0.20	[0.57, 0.67]
	Sindhi	0.64	0.09	[0.64, 0.65]	0.65	0.07	[0.64, 0.66]	0.64	0.07	[0.63, 0.66]
Sentence syntax similarity (adjacent sentences)	Finnish	0.10	0.08	[0.09, 0.12]	0.08	0.05	[0.08, 0.10]	0.07	0.04	[0.07, 0.09]
	Sindhi	0.12	0.06	[0.12, 0.13]	0.12	0.06	[0.11, 0.13]	0.11	0.05	[0.10, 0.13]

Table 12 Supplementary Data: Characteristics of CEFR levels in Finnish and Sindhi: Means, standard deviations, and confidence intervals for phrasal density indices

Syntactic complexity indices		A1			A2			B1		
	Language	M	S.D	[95% CI]	M	S.D	[95% CI]	M	S.D	[95% CI]
Noun phrase density	Finnish	387.38	62.90	[374.41, 400.36]	375.24	55.27	[365.09, 385.40]	346.52	39.79	[337.38, 355.66]
	Sindhi	367.67	57.84	[362.64, 372.71]	365.15	50.12	[359.32, 370.98]	367.85	41.13	[357.29, 378.42]
Verb phrase density	Finnish	229.23	77.22	[214.41, 244.04]	240.31	52.98	[228.72, 251.91]	253.08	51.15	[242.65, 263.52]
	Sindhi	204.27	51.72	[199.74, 208.81]	210.59	44.29	[205.34, 215.85]	220.98	42.54	[211.45, 230.50]
Adverbial phrase density	Finnish	22.37	33.80	[15.87, 28.88]	28.72	23.97	[23.63, 33.81]	31.22	21.92	[26.64, 35.80]
	Sindhi	29.28	24.65	[27.04, 31.51]	30.63	23.83	[28.04, 33.22]	30.47	18.30	[25.78, 35.17]
Preposition phrase density	Finnish	91.70	43.25	[82.69, 100.70]	94.73	32.79	[87.68, 101.77]	94.03	33.12	[87.69, 100.37]
	Sindhi	111.16	42.83	[107.51, 114.80]	120.31	35.23	[116.09, 124.53]	122.71	27.71	[115.06, 130.36]

Negation density	Finnish	58.09	36.51	[51.13, 65.04]	50.61	26.46	[45.17, 56.05]	43.08	22.39	[38.18,47.98]
	Sindhi	51.01	29.18	[48.55, 53.47]	44.21	23.00	[41.37, 47.06]	33.93	19.99	[28.77, 39.09]
Gerund density	Finnish	7.22	19.72	[2.84, 11.61]	8.91	15.26	[5.48, 12.34]	14.97	17.45	[11.88, 18.05]
	Sindhi	10.93	17.08	[9.27, 12.58]	14.67	17.33	[12.75, 16.59]	17.08	19.87	[13.61, 20.56]
Infinitive density	Finnish	6.59	18.43	[1.93, 11.24]	13.95	18.93	[10.31, 17.59]	20.63	17.74	[17.35, 23.91]
	Sindhi	5.44	10.16	[4.40, 6.49]	8.70	11.55	[7.50, 9.91]	11.23	12.85	[9.05, 13.42]

Table 13 Supplementary Data: MANOVA analyses of the differences between Finnish and Sindhi EFL learners at different CEFR levels with respect to syntactic complexity

	Finnish *N*	Sindhi *N*	Wilks's Lambda	*F*	*p*	η^2
A1	54	410	.531	13.20	<.001	.469
A2	82	309	.471	14.00	<.001	.529
B1	110	97	.234	20.00	<.001	.766

N = number of texts in each category (country / language by CEFR level)

Table 14 Supplementary Data: Syntactic complexity indices in EFL writing that differed significantly between Finnish L1 and Sindhi L1 speakers at CEFR level A1

Variable	Language	M	S.D	F	p	d
Mean length of sentence	Finnish	8.10	2.90	59.44	< .001	1.228
	Sindhi	12.24	3.80			
Sentence length (st. dev.)	Finnish	3.12	2.16	41.94	< .001	1.089
	Sindhi	6.32	3.55			
Mean length of T-Unit	Finnish	10.00	3.32	50.72	< .001	1.078
	Sindhi	13.82	3.75			
Mean length of clause	Finnish	7.30	2.33	31.94	< .001	0.928
	Sindhi	10.04	3.46			
T-unit per sentence	Finnish	0.84	0.18	26.42	< .001	0.871
	Sindhi	1.04	0.27			
Complex nominals per clause	Finnish	0.66	0.34	15.768	< .001	0.685
	Sindhi	0.90	0.36			
Left embeddedness (words before main verb)	Finnish	1.63	1.39	19.96	< .001	0.684
	Sindhi	2.66	1.61			
Minimal edit distance (parts of speech)	Finnish	0.47	0.35	45.01	< .001	0.616
	Sindhi	0.63	0.11			
Complex nominals per T-unit	Finnish	1.08	0.49	18.914	< .001	0.584
	Sindhi	1.40	0.60			
Preposition phrase density	Finnish	88.34	45.92	13.98	< .001	0.511

	Sindhi	111.07	41.44			
Clauses per sentence	Finnish	1.60	0.56	12.330	< .001	0.070
	Sindhi	1.64	0.57			
Adverbial phrase density	Finnish	30.21	23.03	9.465	0.002	0.023
	Sindhi	30.76	23.81			
Noun phrase density	Finnish	372.05	44.75	8.484	0.004	0.018
	Sindhi	366.25	50.45			
Negation density	Finnish	49.98	23.90	5.328	0.021	0.011
	Sindhi	44.18	23.14			
Verb phrase density	Finnish	236.33	47.80	4.050	0.045	0.009
	Sindhi	210.66	44.72			
Modifiers per noun phrase	Finnish	0.47	0.16	3.717	0.054	0.008
	Sindhi	0.60	0.20			

Table 15 Supplementary Data: Syntactic complexity indices in EFL writing that differed significantly between Finnish L1 and Sindhi L1 speakers at CEFR level A2

Syntactic Complexity Variables	Language	M	S.D	*F*	*p.*	d
Mean Length of Clause	Finnish	7.51	2.19	61.79	< .001	0.986
	Sindhi	9.72	2.29			
Preposition phrase density	Finnish	92.66	28.59	44.20	< .001	0.871
	Sindhi	120.44	34.84			
Mean Length of T-Unit	Finnish	12.00	3.01	39.29	< .001	0.837
	Sindhi	14.95	3.97			
Complex nominals per clause	Finnish	0.66	0.34	29.72	< .001	0.685
	Sindhi	0.90	0.36			
Number of modifiers per noun phrase	Finnish	0.47	0.16	27.26	< .001	0.717
	Sindhi	0.60	0.20			
Mean length of Sentence	Finnish	11.31	3.11	22.25	< .001	0.623
	Sindhi	13.50	3.87			
Verb phrase density	Finnish	236.33	47.80	20.73	<.001	0.554
	Sindhi	210.66	44.72			
Complex nominals per T-unit	Finnish	1.08	0.49	19.81	<.001	0.584
	Sindhi	1.40	0.60			
Left embeddedness (words before main verb)	Finnish	2.09	1.45	19.10	<.001	0.541
	Sindhi	2.88	1.47			
Coordinate phrases per T-unit	Finnish	0.17	0.25	15.13	<.001	0.489
	Sindhi	0.30	0.28			
Coordinate phrases per clause	Finnish	0.12	0.17	13.37	<.001	0.443
	Sindhi	0.20	0.19			
Gerund density	Finnish	8.06	13.94	9.51	0.002	0.406
	Sindhi	14.46	17.34			
Infinitive density	Finnish	13.72	17.43	11.37	<.001	0.364
	Sindhi	8.38	11.18			
Dependant clauses per T-unit	Finnish	0.63	0.41	10.82	<.001	0.393
	Sindhi	0.48	0.35			
Syntactic simplicity (Z-score)	Finnish	-0.02	1.00	9.51	0.002	0.383
	Sindhi	0.36	0.98			
Sentence length, (S.D)	Finnish	5.91	2.92	8.56	0.004	0.381
	Sindhi	7.14	3.50			
Dependant clauses per clause	Finnish	0.34	0.17	9.13	0.003	0.374
	Sindhi	0.28	0.15			
Syntactic simplicity, percentile	Finnish	50.02	28.12	8.32	0.004	0.356
	Sindhi	59.97	27.67			
Sentence syntax similarity (adjacent sentences)	Finnish	0.10	0.05	7.40	0.007	0.362
	Sindhi	0.12	0.06			
T-units per sentence	Finnish	0.96	0.23	7.23	0.007	0.346
	Sindhi	1.04	0.23			
Complex T-units per t-unit	Finnish	0.44	0.28	5.37	0.021	0.273
	Sindhi	0.37	0.23			
Negation density	Finnish	49.98	23.90	4.01	0.046	0.246
	Sindhi	44.18	23.14			
Verb phrases per T-unit	Finnish	1.88	0.55	3.58	0.059	0.234
	Sindhi	1.76	0.47			

Table 16 Supplementary Data: Syntactic complexity indices in EFL writing that differed significantly between Finnish L1 and Sindhi L1 speakers at CEFR level B1

Variable	Language	M	S.D	*F*	*p*	d
Preposition phrase density	Finnish	91.64	32.01	54.97	< .001	1.03
	Sindhi	122.76	27.85			
Mean length of clause	Finnish	8.27	2.09	35.19	< .001	0.828
	Sindhi	9.94	1.94			
Verb phrase density	Finnish	254.63	46.48	31.46	< .001	0.784
	Sindhi	220.02	41.68			
Dependent clauses per clause	Finnish	0.43	0.15	30.74	< .001	0.758
	Sindhi	0.32	0.14			
Sentence syntax similarity (adjacent sentences)	Finnish	0.08	0.04	26.89	< .001	0.784
	Sindhi	0.12	0.06			
Complex T-unit per t-unit	Finnish	0.61	0.28	26.75	< .001	0.741
	Sindhi	0.42	0.23			
Left embeddedness (words before main verb)	Finnish	2.17	1.36	24.30	< .001	0.682
	Sindhi	3.18	1.59			
Infinitive density	Finnish	21.23	17.25	23.45	< .001	0.680
	Sindhi	10.95	12.59			
Dependant clauses per T-unit	Finnish	0.84	0.48	22.88	< .001	0.679
	Sindhi	0.56	0.33			
Verb phrases per T-unit	Finnish	2.30	0.71	21.83	< .001	0.647
	Sindhi	1.91	0.47			
Noun phrase density	Finnish	345.41	36.82	19.69	< .001	0.616
	Sindhi	369.02	39.69			
Complex nominals per clause	Finnish	0.79	0.33	15.89	< .001	0.529
	Sindhi	0.97	0.35			

Negation density	Finnish	43.61	22.72	11.05	< .001	0.465
	Sindhi	33.67	19.93			
Mean length of T-Unit	Finnish	14.50	3.72	10.79	< .001	0.457
	Sindhi	16.22	3.80			
Coordinate phrases per clause	Finnish	0.13	0.14	8.71	0.004	0.344
	Sindhi	0.18	0.15			
Modifiers per noun phrase	Finnish	0.54	0.21	8.86	0.003	0.460
	Sindhi	0.63	0.18			
Syntactic simplicity, percentile	Finnish	45.75	27.50	7.06	0.009	0.371
	Sindhi	55.60	25.55			
T-units per sentence	Finnish	0.97	0.20	6.88	0.009	0.371
	Sindhi	1.05	0.23			
Syntactic simplicity, z-score	Finnish	-0.14	0.93	6.39	0.012	0.359
	Sindhi	0.18	0.85			
Clauses per T-unit	Finnish	1.83	0.55	5.83	0.017	0.332
	Sindhi	1.67	0.40			
Coordinate phrases per T-unit	Finnish	0.22	0.24	5.21	0.023	0.319
	Sindhi	0.30	0.26			
Complex nominals per T-unit	Finnish	1.39	0.62	4.86	0.029	0.313
	Sindhi	1.58	0.59			
Mean sentence length	Finnish	13.86	3.51	4.18	0.042	0.282
	Sindhi	14.97	4.30			

Table 17 Supplementary Data: Comparison of the mean values for dependent clauses per clause and clauses per T-unit among Sindhi and Finnish L1 learners of English (rated on CEFR levels), and Chinese and Spanish L1 learners of English (rated on IELTS bands; see Banerjee et al. 2007)

	Sindhi L1 (CEFR level)	Finnish L1 (CEFR level)	Chinese L1 (IELTS Band)	Spanish L1 (IELTS Band)
Dependent clauses per clause				.59 (8) .67 (7)
			.47 (6)	.47 (6)
	.32 (B1)	.43 (B1)	.41 (5)	.61 (5) .42 (4)
	.28 (A2)	.34 (A2)	.35 (4)	
	.25 (A1)	.27 (A1)	.27 (3)	
Clauses per T-unit				2.55 (8) 3.17 (7)
	1.66 (B1)	1.83 (B1)	1.89 (6) 1.71(5)	1.96 (6) 2.59 (5)
	1.58 (A2)	1.71 (A2)	1.54 (4)	1.76 (4)
	1.46 (A1)	1.49 (A1)	1.54 (4) 1.43 (3)	

Table 18 Supplementary Data: Summary of SC indices that separated between adjacent CEFR levels among Sindhi and Finnish L1 learners of English

A1 vs A2 **Finnish**	A1 vs A2 **Sindhi**	A2 vs B1 **Finnish**	A2 vs B1 **Sindhi**
Mean length of sentence	**Mean length of sentence**	**Mean length of sentence**	*Mean length of sentence*
Sentence length (st. dev.)	**Sentence length (st. dev.)**	**Sentence length (st.dev.)**	
	Mean length of T-Unit	**Mean length of T-Unit**	*Mean length of T-Unit*
Verb phrases per T-unit	**Verb phrases per T-unit**	**Verb phrases per T-unit**	*Verb phrases per T-unit*
Clauses per sentence	*Clauses per sentence*		
	Dependant clauses per clause	*Dependant clauses per clause*	*Dependant clauses per clause*
T-units per sentence			
	Complex T-units per T-unit	*Complex T-units per T-unit*	*Complex nominals per T-unit*
	Clauses per T-unit		
	Dependent clauses per T-unit		
Minimal edit distance			
Modifiers per noun phrase	*Left embeddedness*	*Modifiers per noun phrase*	
Infinitive density	**Infinitive density**	**Noun phrase density**	
	Negation density		**Negation density**
	Preposition phrase density	Infinitive density	
	Gerund density	Gerund density	

SC indices in **bold** font refer to indices that distinguished CEFR levels at $p < .001$; the indices in *italic* font refer to $.01 < p < 001$) and those with regular font to $.05 < p < .01$.

Table 19 Supplementary Data: Summary of SC indices that separated between the non-adjacent CEFR levels A1 and B1 among Sindhi and Finnish L1 learners of English

A1 vs B1	
Finnish	**Sindhi**
Mean length of sentence	**Mean length of sentence**
Sentence length (st. dev)	**Sentence length (st. dev.)**
Mean length of T-Unit	**Mean length of T-Unit**
Mean length of clause	
Clauses per sentence	*Clauses per sentence*
T-units per sentence	
Clauses per T-unit	**Clauses per T-unit**
Verb phrases per T-unit	**Verb phrases per T-unit**
Dependant clauses per clause	**Dependant clauses per clause**
Dependant clauses per T-unit	**Dependent clauses per T-unit**
Complex T-units per T-unit	**Complex T-units per T-unit**
Complex nominals per T-unit	**Complex nominals per T-unit**
Minimal edit distance	**Left embeddedness**
Infinitive density	**Infinitive density**
Noun phrase density	
Negation density	**Negation density**
Gerund density	*Gerund density*
	Verb phrase density
	Preposition phrase density

The SC indices in **bold** font refer to indices that distinguished CEFR levels at $p < .001$; the indices in *italic* font refer to $.01 < p < 001$).

Appendix 1 Supporting Data: The writing task

The box below presents the task and task instruction given to the learners. In Pakistan, the task and instruction were translated orally by the researcher into Sindhi (the first language of most of the students) and Urdu (the language that most Pakistanis understand irrespective of their first language). In Finland, the task instruction was also given orally in Finnish.

Write on the following topic; what you think about the matter. Give reasons for your opinion.

Topic: No mobile phones at school! (Give at least a few reasons)

Write in English in clear characters in the space below (continues on the reverse side).

Write at least 50 words

Appendix 2 Supporting Data: Analyses of the ratings

The following four pictures present the main stages in the analysis of the ratings of the data collected from Sindhi-speaking EFL writers. The pictures are screenshots of the Rater Measurement Reports from Facets analyses and display information about raters' consistency (Infit and Outfit Mean squares and standardized fit statistics, ZStd) and leniency/severity (values in the Measure column, and more transparently, in the Observed and Fair Average columns).

T1 below presents the analysis in which all raters were included. Rater100 and Rater200 were Finnish raters who had been involved in the rating of the Finnish EFL writing performance and who rated a proportion of the performances collected from the Sindhi-speaking EFL learners. All other raters were Pakistani raters who were trained but who had not previously rated performances against the CEFR levels.

T1 shows that Rater4 differs from the others in two important ways: he/she is quite unsystematic as shown by high Infit and Outfit mean square values (1.50 and 1.44, respectively) which means that he/she sometimes rated unexpectedly harshly and sometimes unexpectedly leniently compared to the other raters who had rated the same performances and also with respect to his/her own level of leniency/severity. Rater4 is also the most severe rater of all: the logit measure for that rater is -4.38 which is clearly lower than for the next most severe rater (Rater9). Therefore, Rater 4 was removed from the data.

T1

Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Num	Rater
624	189	3.3	3.27	4.68	.12	.75	-2.6	.95	-.4	1.07	.55	5.8	9.1	6	Rater6
498	252	2.0	1.91	1.60	.11	.63	-4.5	.70	-3.4	1.30	.66	30.3	35.6	100	Rater100
476	237	2.0	1.88	1.52	.11	1.35	3.2	1.15	1.3	.78	.36	37.6	26.9	2	Rater2
274	180	1.7	1.86	1.47	.14	.97	-.2	1.07	.6	.93	.64	45.8	53.5	23	Rater23
272	180	1.6	1.85	1.43	.14	.95	-.3	1.09	.7	.86	.71	50.6	53.6	26	Rater26
377	189	2.0	1.83	1.36	.12	.87	-1.2	.96	-.2	1.01	.74	25.9	28.8	8	Rater8
342	178	2.0	1.72	1.02	.13	.92	-.6	.84	-1.3	1.11	.60	43.3	43.5	15	Rater15
99	60	1.7	1.63	.72	.24	.49	-3.2	.48	-3.1	1.52	.76	41.4	43.0	200	Rater200
229	180	1.4	1.56	.50	.15	1.07	.6	1.02	.1	.93	.72	42.8	52.5	25	Rater25
308	179	1.8	1.52	.36	.14	.67	-3.1	.72	-2.5	1.29	.63	44.0	47.1	11	Rater11
301	237	1.3	1.14	-1.31	.14	1.24	2.1	.97	-.1	.98	.27	44.7	40.0	3	Rater3
203	147	1.4	1.12	-1.41	.17	1.06	.4	.74	-1.6	1.08	.30	39.9	53.9	17	Rater17
216	179	1.2	1.09	-1.67	.16	.88	-.9	.88	-.8	1.14	.51	33.0	38.4	13	Rater13
221	189	1.2	1.09	-1.68	.16	1.55	3.8	1.09	.6	.92	.52	31.9	32.9	7	Rater7
193	147	1.3	1.08	-1.70	.17	1.12	.9	.75	-1.5	1.14	.20	37.6	53.5	18	Rater18
193	189	1.0	1.00	-2.51	.18	1.02	.1	1.06	.3	.96	.55	30.3	29.8	9	Rater9
160	237	.7	.74	-4.38	.16	1.50	4.7	1.44	3.0	.52	.31	31.1	26.4	4	Rater4
293.3	185.2	1.6	1.54	.00	.15	1.00	.0	.94	-.5		.53				Mean (Count: 17)
131.6	43.0	.6	.57	2.07	.03	.28	2.5	.22	1.6		.17				S.D. (Population)
135.6	44.3	.6	.58	2.13	.03	.29	2.6	.22	1.7		.18				S.D. (Sample)

Model, Populn: RMSE .15 Adj (True) S.D. 2.06 Separation 13.44 Strata 18.25 Reliability (not inter-rater) .99
Model, Sample: RMSE .15 Adj (True) S.D. 2.12 Separation 13.86 Strata 18.81 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 3719.9 d.f.: 16 significance (probability): .00
Model, Random (normal) chi-square: 15.9 d.f.: 15 significance (probability): .39
Inter-Rater agreement opportunities: 4013 Exact agreements: 1361 = 33.9% Expected: 1439.9 = 35.9%

T 2 below displays the same Rater Measurement Report after the rating data were re-analysed without Rater4. At this stage, we decided to remove Rater6. That rater worked very consistently as indicated by the perfectly acceptable Infit and Outfit values (mean squares are close to the optimal 1.0, and in general, lower than 1 values are not problematic in the same way as values that exceed 1.0 considerably). However, Rater6 clearly has a very different understanding of the CEFR levels since he/she is much more lenient than the other raters. Both observed and adjusted fair average values for that rater were well over 3 (i.e., CEFR level B1) whereas for the next most lenient raters,

these values were below 2 (i.e., CEFR level A2). In order for the ratings of the Sindhi learners (particularly those that Rater6 had assessed) not to be inflated because of the (too) lenient judgments by Rater6, his/her ratings were removed from the data.

T2

Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Num	Rater
624	189	3.3	3.26	4.73	.12	.81	-1.9	1.05	.4	1.00	.55	5.8	8.8	6	Rater6
498	252	2.0	1.90	1.42	.11	.67	-3.9	.75	-2.7	1.26	.66	31.4	37.1	100	Rater100
274	180	1.7	1.86	1.28	.15	1.04	.3	1.16	1.2	.84	.64	45.8	54.9	23	Rater23
476	237	2.0	1.85	1.24	.12	.99	.0	.85	-1.4	1.08	.27	48.0	42.2	2	Rater2
272	180	1.6	1.85	1.24	.15	1.02	.2	1.20	1.4	.77	.71	50.6	55.1	26	Rater26
377	189	2.0	1.83	1.18	.13	.94	-.5	1.06	.5	.93	.74	25.9	28.8	8	Rater8
342	178	2.0	1.74	.86	.14	.99	.0	.90	-.8	1.04	.60	43.3	44.6	15	Rater15
99	60	1.7	1.63	.49	.25	.52	-2.9	.50	-2.8	1.49	.76	41.4	43.9	200	Rater200
229	180	1.4	1.55	.24	.16	1.14	1.2	1.09	.7	.85	.72	42.8	53.7	25	Rater25
308	179	1.8	1.53	.15	.14	.72	-2.6	.78	-1.8	1.24	.63	44.0	48.4	11	Rater11
203	147	1.4	1.12	-1.73	.17	1.13	1.0	.79	-1.2	1.02	.30	39.9	55.5	17	Rater17
301	237	1.3	1.10	-1.92	.14	1.15	1.3	.97	-.1	1.01	.26	48.7	39.9	3	Rater3
193	147	1.3	1.09	-2.04	.18	1.21	1.6	.79	-1.1	1.07	.20	37.6	55.0	18	Rater18
216	179	1.2	1.09	-2.04	.17	.95	-.4	.97	-.1	1.07	.51	33.0	38.8	13	Rater13
221	189	1.2	1.08	-2.09	.17	1.68	4.6	1.21	1.2	.82	.52	31.9	33.3	7	Rater7
193	189	1.0	1.00	-3.00	.19	1.12	.9	1.22	1.0	.88	.55	30.3	30.0	9	Rater9
301.6	182.0	1.7	1.59	.00	.16	1.01	-.1	.95	-.4		.54			Mean (Count: 16)	
131.2	42.2	.5	.54	1.94	.03	.25	2.0	.20	1.4		.18			S.D. (Population)	
135.5	43.6	.5	.56	2.00	.03	.26	2.1	.20	1.4		.19			S.D. (Sample)	

Model, Populn: RMSE .16 Adj (True) S.D. 1.93 Separation 12.15 Strata 16.53 Reliability (not inter-rater) .99
Model, Sample: RMSE .16 Adj (True) S.D. 1.99 Separation 12.55 Strata 17.06 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 2903.9 d.f.: 15 significance (probability): .00
Model, Random (normal) chi-square: 14.9 d.f.: 14 significance (probability): .38
Inter-Rater agreement opportunities: 3505 Exact agreements: 1203 = 34.3% Expected: 1332.1 = 38.0%

T3 below displays the Rater Measurement Report after both Rater4 and Rater6 were removed from the analysis. We can see that there is still variation in rater leniency/severity but that has been considerably reduced and there is no rater who clearly differs from the others in this respect (Fair average values – values that Facets adjusts, to some extent, to account for rater leniency and learners' abilities – range now from 1.01 to 1.85). However, Rater7 stands out as being quite inconsistent in his/her ratings (e.g. Infit Mean square is 1.76). Therefore, that rater's ratings were removed from the analysis.

T3

Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Num	Rater
498	252	2.0	1.85	1.73	.11	.76	-2.6	.88	-1.1	1.20	.64	33.6	41.1	100	Rater100
377	189	2.0	1.83	1.66	.13	.70	-2.8	.88	-.9	1.16	.63	31.4	34.8	8	Rater8
274	180	1.7	1.82	1.64	.15	1.05	.4	1.26	1.6	.79	.64	45.8	56.1	23	Rater23
272	180	1.6	1.81	1.60	.15	1.00	.0	1.33	2.0	.67	.71	50.6	56.2	26	Rater26
476	237	2.0	1.78	1.49	.12	.84	-1.4	.69	-2.8	1.20	.27	48.0	44.9	2	Rater2
342	178	2.0	1.68	1.14	.13	.93	-.5	.85	-1.1	1.09	.60	43.3	45.8	15	Rater15
99	60	1.7	1.63	.94	.25	.62	-2.1	.58	-1.9	1.41	.67	47.4	51.1	200	Rater200
229	180	1.4	1.52	.59	.16	1.21	1.7	1.22	1.4	.72	.72	42.8	54.9	25	Rater25
308	179	1.8	1.48	.43	.14	.73	-2.3	.83	-1.1	1.22	.63	44.0	49.6	11	Rater11
203	147	1.4	1.11	-1.45	.18	1.19	1.3	.81	-.9	.96	.30	39.9	57.0	17	Rater17
221	189	1.2	1.08	-1.71	.18	1.76	4.8	1.15	.8	.86	.57	41.5	47.1	7	Rater7
301	237	1.3	1.08	-1.73	.15	1.14	1.1	.96	-.2	1.02	.26	48.7	42.7	3	Rater3
193	147	1.3	1.08	-1.77	.18	1.29	2.0	.80	-.9	.99	.20	37.6	56.6	18	Rater18
216	179	1.2	1.07	-1.85	.18	1.01	.0	1.03	.2	1.01	.51	33.0	39.8	13	Rater13
193	189	1.0	1.01	-2.72	.20	1.18	1.2	1.73	2.6	.83	.55	40.3	43.0	9	Rater9
280.1	181.5	1.6	1.46	.00	.16	1.03	.1	1.00	-.2		.53			Mean (Count: 15)	
104.7	43.6	.3	.33	1.59	.03	.28	2.0	.28	1.5		.17			S.D. (Population)	
108.4	45.1	.3	.34	1.65	.04	.29	2.1	.29	1.6		.18			S.D. (Sample)	

Model, Populn: RMSE .16 Adj (True) S.D. 1.58 Separation 9.68 Strata 13.25 Reliability (not inter-rater) .99
Model, Sample: RMSE .16 Adj (True) S.D. 1.64 Separation 10.03 Strata 13.70 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 1510.6 d.f.: 14 significance (probability): .00
Model, Random (normal) chi-square: 13.9 d.f.: 13 significance (probability): .38
Inter-Rater agreement opportunities: 2807 Exact agreements: 1151 = 41.0% Expected: 1312.7 = 46.8%

T4 below displays the final Rater Measurement Report. Raters 4, 6 and 7 were removed from the analysis, as well as about 20 individual data points (individual ratings) that the analysis indicated as being much unexpected given the particular rater's overall leniency/severity and what the other raters had given to that particular performance. All the Infit and Outfit values are reasonably good (mean squares are smaller than 1.50; for a discussion of the meaning and magnitude of these values in Rasch measurement, see e.g. https://www.rasch.org/rmt/rmt162f.htm). Ideally, all the raters would be very similar in terms of leniency/severity, that is, they should have a very similar understanding of the meaning of the rating scale (CEFR levels in this case). In the current data, the raters can be roughly divided into those whose average rating are closer to A1 (Fair-M Averages close to 1.0) and those whose ratings are closer to A2 (Fair-M values close to 2.0). However, all average ratings are within one CEFR level (their spread is less than 1.0 on the Fair average / rating scale). In practice, the fact that each performance was multiply rated reduces the likelihood that particular performances would be 'misplaced' on the CEFR scale because they happened to be rated only by the most severe vs the most lenient raters. At the end of the day, however, no performance can be placed on a particular level of the rating scale with 100% certainty. Therefore, a reasonably large number of performances are needed if one is interested in analyzing the characteristics of the CEFR or other proficiency scale levels.

T4

Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Num	Rater
498	252	2.0	1.88	1.76	.11	.61	-4.5	.76	-2.3	1.31	.68	36.5	42.8	100	Rater100
274	180	1.7	1.85	1.65	.15	1.12	.9	1.32	1.9	.74	.64	46.0	57.4	23	Rater23
272	180	1.6	1.84	1.60	.15	1.07	.6	1.41	2.4	.60	.72	50.8	57.5	26	Rater26
377	189	2.0	1.83	1.55	.13	.59	-4.0	.70	-2.5	1.28	.54	34.8	35.6	8	Rater8
476	237	2.0	1.83	1.54	.12	.77	-2.2	.73	-2.4	1.22	.32	48.1	44.0	2	Rater2
342	178	2.0	1.72	1.12	.14	1.02	.1	.91	-.6	1.01	.60	43.3	46.6	15	Rater15
99	60	1.7	1.67	.94	.26	.58	-2.3	.54	-2.2	1.43	.71	50.6	53.2	200	Rater200
225	179	1.4	1.54	.46	.17	1.22	1.8	1.23	1.5	.69	.73	43.1	55.6	25	Rater25
308	179	1.8	1.51	.34	.15	.79	-1.7	.89	-.7	1.17	.63	44.0	50.6	11	Rater11
203	147	1.4	1.11	-1.69	.18	1.26	1.8	.85	-.6	.90	.30	39.9	57.7	17	Rater17
193	147	1.3	1.08	-2.03	.19	1.36	2.5	.86	-.6	.91	.20	37.6	57.3	18	Rater18
296	236	1.3	1.08	-2.04	.16	1.00	.0	.95	-.3	1.04	.32	48.9	41.5	3	Rater3
216	179	1.2	1.07	-2.11	.18	1.07	.6	1.14	.8	.94	.51	33.0	39.6	13	Rater13
193	189	1.0	1.01	-3.09	.20	1.21	1.5	1.89	3.0	.75	.47	20.7	24.8	9	Rater9
283.7	180.9	1.6	1.50	.00	.16	.98	-.3	1.01	-.2		.52			Mean (Count: 14)	
107.3	45.0	.3	.34	1.71	.04	.25	2.2	.34	1.8		.17			S.D. (Population)	
111.3	46.7	.3	.35	1.77	.04	.26	2.3	.35	1.9		.18			S.D. (Sample)	

Model, Populn: RMSE .17 Adj (True) S.D. 1.70 Separation 10.15 Strata 13.86 Reliability (not inter-rater) .99
Model, Sample: RMSE .17 Adj (True) S.D. 1.76 Separation 10.53 Strata 14.38 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 1518.9 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.9 d.f.: 12 significance (probability): .38
Inter-Rater agreement opportunities: 2311 Exact agreements: 946 = 40.9% Expected: 1085.1 = 47.0%

II

SYNTACTIC COMPLEXITY IN FINNISH-BACKGROUND EFL LEARNERS' WRITING AT CEFR LEVELS A1 – B2

by

Khushik, G. A., & Huhta, A. 2022.

European Journal of Applied Linguistics, 10(1), 142-184.

https://doi.org/10.1515/eujal-2021-0011

Reproduced with kind permission by Walter de Gruyter.

Syntactic complexity in Finnish-background EFL learners' writing at CEFR levels A1 – B2

Abstract

The increasing importance of the Common European Framework of Reference (CEFR) has led to research on the linguistic characteristics of its levels, as this would help the application of the CEFR in the design of teaching materials, courses, and assessments. This study investigated whether CEFR levels can be distinguished with reference to syntactic complexity (SC). 14- and 17-year-old Finnish learners of English (N=397) wrote three writing tasks which were rated against the CEFR levels. The ratings were analysed with multi-facet Rasch analysis and the texts were analysed with automated tools. Findings suggest that the clearest separators at lower CEFR levels (A1-A2) were the mean sentence and T-unit length, variation in sentence length, infinitive density, clauses per sentence or T-unit, and verb phrases per T-unit. For higher levels (B1-B2) they were modifiers per noun phrase, mean clause length, complex nominals per clause, and left embeddedness. The results support previous findings that the length of and variation in the longer production units (sentences, T-units) are the SC indices that most clearly separate the lower CEFR levels, whereas the higher levels are best distinguished in terms of complexity at the clausal and phrasal levels.

Keywords

English as a foreign language (EFL), Syntactic complexity, Common European Framework of Reference (CEFR), Automated analysis of learners' written scripts

Abstrakti

Eurooppalaisen viitekehyksen (EVK) merkitys kielikoulutukselle on lisännyt tutkimusta sen taitotasojen kielellisistä piirteistä; tarkempi tieto näistä piirteistä auttaisi EVK:n soveltamista opetusmateriaalien, kurssien ja arviointin laatimiseen. Tutkimuksessa selvitettiin eroavatko EVK:n tasot toisistaan syntaksin kompleksisuuden perusteella. Suomalaiset 14- ja 17-vuotiaat englannin oppijat (N=379) kirjoittivat kolme kirjoitelmaa, jotka arvioitiin EVK:n taitotasoille. Arviointiaineisto tutkittiin monitahoisella Rasch-analyysillä ja tekstien piirteet selvitettiin automaattisilla analyysiohjelmilla. Tuloksien perusteella alimpia EVK-tasoja (A1-A2) erotti selvimmin toisistaan lauseiden ja T-yksiköiden pituus, vaihtelu lauseiden pituudessa, infinitiivirakenteiden määrä, lausekkeiden ja T-yksiköiden määrä lauseissa ja verbirakenteiden määrä T-yksiköissä. Ylempiä tasoja (B1-B2) erottelivat puolestaan määritteiden määrä nominifraaseissa, lausekkeiden pituus, kompleksisten rakenteiden määrä lausekkeissa ja pääverbiä edeltävien sanojen määrä (left embeddedness). Tulokset ovat linjassa aiempien syntaksin kompleksisuuden tutkimusten kanssa siinä, että pidempien tuotosyksikköjen (lauseet, T-yksiköt) pituus ja vaihtelu erottelee selvimmin englannin oppijoita alemmilla EVK-tasoilla, kun taas korkeammilla taitotasoilla erot ilmenevät lausekkeiden ja fraasien käytössä.

Avainsanat

Englanti vieraana kielenä, syktaktinen kompleksisuus, Yhteinen Eurooppalainen Viitekehys (EVK), oppijoiden kirjoituksen automaattinen analyyysi

Sammandrag

Den ökande vikten av den allmäneuropeiska referensramen (CEFR) har lett till forskning i lingvistiska egenskaper hos CEFR-nivåerna eftersom den kan främja tillämpandet av CEFR i planeringen av undervisningsmaterial, kurser och bedömning. I denna studie undersöktes det om det finns skillnader i syntaktisk komplexitet (SK) mellan CEFR-nivåerna. 14- och 17-åriga finskspråkiga studerande av engelska (N=397) skrev tre skrivuppgifter som bedömdes enligt CEFR-nivåerna. Bedömningarna analyserades med mångfasetterad Rasch-analys och texterna analyserades med automatiserade verktyg. Fynden tyder på att de tydligaste särskiljande faktorerna på de lägre CEFR-nivåerna (A1-A2) var den genomsnittliga längden på meningar och T-enheter, variationen i meningslängden, tätheten av infinitiver, antalet satser per mening eller T-enhet och antalet verbfraser per T-enhet. På de högre nivåerna (B1-B2) var faktorerna antal bestämningar per nominalfras, genomsnittlig satslängd, antal komplexa nominala per sats och antal ord före huvudverb (left embeddedness). Resultaten stöder tidigare fynden om att längden på och variationen i längre produktionsenheter (meningar, T-enheter) är de SK tecken som tydligaste gör skillnader mellan de lägre CEFR-nivåerna, medan de högre nivåerna skiljer sig mest från varandra i komplexitet på sats- och frasnivåerna.
Nyckelord

Engelska som främmande språk, syntaktisk komplexitet, allmäneuropeiska referensramen, automatisk analys av elevers texter

1 Introduction

The Common European Framework of Reference (CEFR; Council of Europe 2001) is arguably the most influential initiative in foreign language education from Europe. Since its introduction, the CEFR has rapidly become *the* framework for language education across Europe. The CEFR is seen to have general value for language learning, teaching and assessment. Mainly its 6-point scale defining degrees of proficiency from basic to very advanced is now widely used to describe the level of language examinations, curricula, courses, materials, and targets for learning. The importance of the CEFR has, however, brought attention to its limitations.

The most severe issue with the CEFR is probably that its proficiency scale (or its 50+ scales, in fact) is not adequately informed by second language acquisition (SLA) research (Hulstijn 2007, Hulstijn et al. 2010, North 2007, Wiśniewski 2017), even if the scale appears to define developmental stages in learning. A related limitation of the CEFR levels for applying them to the design of level-specific materials, curricula, and assessments is that they define what learners can do with the language; they do not specify which linguistic characteristics (e.g., words and structures) are required, or typically used, in particular foreign languages to the functions and activities described at each level.

These issues have led to calls for research on the relationship between the framework levels and the development of the linguistic aspects of proficiency. Language testers have been at the forefront of applying the CEFR and have faced the framework's limitations (e.g., Alderson, 2007). To increase the validity and applicability of the CEFR levels, language testers and SLA researchers have conducted (often) joint research on the linguistic characteristics of the CEFR levels (see Bartning et al. 2010 and the studies reviewed below). Particularly the language testers interested in diagnostic assessment, that is, predicting and understanding learners' strengths and weaknesses in their L2 skills in order to provide feedback to learners and propose action to address the identified weaknesses, have promoted such research (see Alderson, 2007; Bartning et al., 2010; Author2 et al., forthcoming).

Such collaboration has many benefits. SLA researchers can use the CEFR levels as a reference point, which improves the interpretability of the findings as such levels define informants' second or foreign language (L2) proficiency more transparently than in many previous SLA studies (Carlsen 2012). For their part, language testers can improve the validity of their assessments by grounding them better in SLA research.

The current study contributes to ongoing research on the linguistic basis of the CEFR by investigating two groups of teenage (14 and 17-year-old) Finnish-speaking learners of English as a foreign language (EFL). The study focuses on syntactic complexity (SC) in the learners' writing: how SC relates to communicative CEFR levels (i.e., writing ability as defined in those levels), and whether particular levels can be distinguished from one another in terms of SC.

2 Syntactic complexity in relation to CEFR levels

Syntactic complexity (SC) has been defined variously in the literature. In SLA research, the T-unit has been a critical index in SC analyses (e.g., Wolfe-Quintero et al. 1998), but several other indices have also been investigated, such as mean length of clause (e.g., Ortega 2003) or complex phrases and complex nominals per clause or T-unit (e.g., Lu 2011; for reviews, see, e.g. Wolfe-Quintero et al.1998 and Ortega 2003). Language testers

investigating the linguistic characteristics of different proficiency levels have used the same SC indices as SLA researchers (e.g., Lu 2011, Kyle and Crossley 2017).

Irrespective of how SC is defined and operationalised, it should be seen as part of a system that comprises several levels and dimensions. Bulté and Housen (2012, 2014) argue that SC is part of linguistic complexity, which, in turn, is part of absolute complexity that concerns the number of different components of a particular linguistic feature and the relationships between those components. SC, Bulté and Housen (2012) maintain, comprises three levels: theoretical (number of syntactic structures and their relationships), observational (how different language forms contribute to complexity at the sentence, clause, and phrase levels), and operational (quantitative indices of SC). Our study agrees with Bulté and Housen's (2014: 45–46) definition of complexity "as an absolute, objective, and essentially quantitative property of language units, features, and (sub)systems thereof in terms of (i) the number and the nature of discrete parts that the unit/feature/system consists of and (ii) the number and the nature of the interconnections between the parts".

It should be mentioned that conceptualising SC in terms of the indices of complexity typically used in SLA and some language testing research (e.g., mean length of T-units) are rather broad and have their limitations. Biber et al. (2020) argue that such omnibus measures are pretty extensive in linguistic terms and, thus, not easy to interpret linguistically, and a more detailed description of the structural, syntactic and functional features of the various linguistic elements are needed. This is an obvious limitation of such indices for attempts to develop diagnostic tests even if the broad indices of complexity may suffice for the prediction stage in diagnostic assessment (e.g., Author2 et al., forthcoming). Furthermore, findings from Multidimensional studies on register variation in speaking and writing indicate that grammatical complexity features often vary from one register to another (e.g. Biber, 1992; Biber et al., 2020). Thus, findings from different studies may vary due to the different registers that the writers used to elicit.

Since the current study is part of language testing research that aims to predict L2 learners proficiency level from syntactic complexity in their writing, we use traditional omnibus indices of SC. We also use data based on learner performances across several writing tasks, even though that unavoidably hides possible variation in SC due to some register differences (see the Methods section for more information about the tasks).

Next, we review the literature on the relationship between SC in written L2 English and the CEFR proficiency levels. An early study by Kim (2004) explored SC in 33 scripts rated on CEFR scales. She found some SC features to distinguish levels A2 and B2: adverbial and adjective clauses per clause, clauses and dependent clauses per T-unit, dependent clauses per clause, and prepositional, participial and infinitive phrases per clause.

Hawkins and Filipović (2012) and Green (2012) explored the CEFR-related Cambridge Learner Corpus and found that mean sentence length significantly differentiated all adjacent levels from A2 to C2. In addition, Green (2012) found the mean noun phrase incidence and the mean number of modifiers per noun to differentiate B2 and C1, and sentence syntax similarity to distinguish C1 from C2.

Verspoor et al. (2012) explored descriptive texts written by teenaged Dutch EFL learners on different topics and rated on a 5-point scale corresponding to CEFR levels A1.1, A1.2, A2, B1.1, and B1.2. They found that simple versus complex sentences were strong proficiency level differentiators. Furthermore, sentence length differentiated the proficiency levels and that T-unit length increased from low to high proficiency levels, significantly differentiating A1.2 versus B1.1 and A2 versus B1.2. Relative clauses also

increased across levels showing apparent differences between A2 and B1.1. The number of dependent clauses proved to be the only SC feature that differentiated across all adjacent levels studied. Gyllstad et al. (2014) analysed emails and stories written by 54 L1 Swedish EFL learners who were rated to represent CEFR levels A (A1-A2) or B (B1-B2). The researchers found the mean length of T-units, mean length of clauses, and clauses per T-unit to differentiate between A and B levels.

Alexopoulou et al. (2017) explored SC in EFL learners' texts, analysing the EFCAMDAT Corpus (http://corpus.mml.cam.ac.uk/efcamdat) based on learners from different L1 backgrounds. They reported an increase in sentence length (across all CEFR levels), clause length (from A2 to B2), and clauses per T-unit (from A1 to B2) but did not report on the statistical significance of their findings. Barrot and Agdeppa (2021) used another corpus (ICNALE-Written; http://language.sakura.ne.jp/icnale/download.html) comprising essays written by EFL learners from 10 Asian countries. Over 5,000 essays placed at A2, B1.1, B1.2 and B2 (or above) based on learners' TOEFL and other EFL test results were investigated for 14 SC indices. They found several indices to distinguish those CEFR levels, particularly length of clauses, sentences and T-units, and complex nominals per clause or T-unit. Martínez (2018) investigated 188 Spanish secondary level EFL learners who wrote on an opinion topic. The students were from two grades corresponding to A2 and B1 levels. Her study used SC indices proposed by Bulté and Housen (2014), which differ somewhat from those used in most CEFR-related SC studies. Martínez reported significant differences in the length of that-sentences, compound and complex sentence ratios, coordinate and dependent clause ratios, and noun phrases per clause. Finally, the Authors (XXXX) compared teenaged EFL learners from two L1 (Finnish and Sindhi) backgrounds. Investigating one argumentative writing task and almost 30 indices of syntactic complexity, they discovered that most indices differentiated CEFR levels from A1 to B1 but that the results varied depending on the learners' L1.

Previous research on SC across CEFR levels is, thus, rather heterogeneous. The studies often focus on only a few and different, indices making it challenging to form an overall picture of which features differentiate CEFR levels in EFL learners' writing. The research methods in previous studies also vary considerably. For example, the number and type of the writing tasks vary, as do the conditions under which the tasks are completed. Furthermore, the small scale of some studies and the uncertain reliability of the placement of the writing samples on the CEFR levels make the specific conclusions uncertain. However, a consistent finding is that many SC indices increase as writing ability (CEFR level) improves.

The present study departs from most previous ones in at least three ways. First, it covers a wide range of SC indices to obtain a comprehensive picture of the relationship. Secondly, learners' writing skills were measured by combining the results of several writing tasks because we investigate the SC typical of *learners' writing at different proficiency levels* rather than *particular tasks* (see Methods section). Thirdly, special attention was paid to the reliable placement of learners' scripts at the CEFR levels through direct double rating on the levels and the use of multi-facet Rasch analysis to mitigate unavoidable rater differences.

3 Methods

3.1 Goal and research questions

The study's goal was to shed light on the linguistic characteristics of the CEFR levels by focusing on syntactic complexity. The research questions were:

RQ1. To what extent is the syntactic complexity in the writing of two age groups of Finnish EFL learners related to their EFL writing ability? Which SC indices correlate strongest with their ability, and do the two age groups differ?

RQ2. Which SC indices distinguish Finnish EFL learners at different CEFR levels, and do the two age groups differ?

To answer the RQs, we draw on a corpus of texts written by teenaged EFL learners collected in a research project focusing on reading and writing development in L1 Finnish and L2 English (Author et al. XXXX). The corpus was collected from volunteer learners who completed the tasks in separate data collection sessions in their schools supervised by researchers. The learners were given feedback on their performance, but the tasks were not used for grading purposes.

3.2 Participants

Participants represent two groups of EFL learners with Finnish as their L1: 14-year-olds in grade 8 in the lower-secondary school (N=202) and 17-year-olds in grade two in the academic upper-secondary school (gymnasium, N=195).

3.3 Tasks

Both groups completed three writing tasks: one shared by both and two unique to the group. The shared task was designed in an earlier project focusing on L2 writing in Finland. The task was to express an opinion on one of two topics (should mobile phones be allowed at school; should boys and girls be integrated into the class) and give reasons for their views. The task was based on considering the national curricula for EFL in secondary education; the researchers (university language teachers and researchers) considered the task to enable the stronger (B1-B2) students to display their writing ability while also the weaker (A1-A2) students could address the topics. The unique tasks came from the Pearson Test of English General (Pearson collaborated with the large scale project): the two 8th graders' tasks were from the PTE B1-level test and the two gymnasium tasks from their B2-level test. The PTE tasks were retired operational tasks developed (including standard-setting to CEFR levels) by Pearson item writers. The B1 tasks were primarily descriptive, whereas the B2 tasks were similar to the shared task as they involved expressing a viewpoint and justifying it. The topics related chiefly to travelling (e.g., B1: travelling preferences between home and school; B2: opinion on cheap air travel; why a particular journey had been so unforgettable). The students were not told how their writing would be rated; they likely thought they would be evaluated the same way their teacher(s) would do – which is known to vary, as teachers have great freedom to implement assessment in the Finnish educational system.

3.4 Ratings and rating analyses

An overlapping rating design was used that allowed the linking of all raters and tasks. Each rater was given a randomised batch of handwritten texts representing several tasks

from both student groups. All texts (3 texts x 397 students; totalling 1180 texts as some students wrote only two texts due to absence from one data collection) were rated by two raters out of a pool of 11 raters. The raters were not told which texts were written by which age group. The raters were trained using the CEFR writing scales, the international benchmarks from the Council of Europe website, and local benchmarks from the earlier writing-focused study. The raters then assessed the texts on the CEFR scale A1-C2. The rating scale was a compilation of several scales taken verbatim from the CEFR, namely overall written production; written interaction; correspondence; notes, messages, forms; creative writing; thematic development; and coherence & cohesion. The scale, thus, focused on the communicative quality of the texts. We excluded the CEFR scales that explicitly address grammatical or lexical aspects of proficiency to decrease potential circularity in the data. Raters can be influenced by other features in learners' writing (e.g., syntactic complexity) than those defined in the scale.

Ratings were coded for analysis by converting CEFR levels ratings to numbers (A1=1, A2=2, B1=3, B2=4). Multi-facet Rasch analysis was then conducted in Facets (Linacre 2009) on the combined 8th and gymnasium rating data, including all tasks and raters. Facets are currently the standard approach to analysing ratings in language testing (e.g., McNamara and Knoch 2012; Aryadoust, Ng and Sayama 2021) as it can adjust differences in rater severity and task difficulty when estimating learner ability to produce an ability measure that is more accurate than, for example, an average across (raw) ratings. Furthermore, the ability measures for learners from Facets are equal-interval scale values (logit values) accompanied by parallel ability measures called fair averages that are on the same metrics as the CEFR based rating scale. Thus, in our study, we categorised the learners onto the levels A1-B2 for investigating whether specific SC indices differentiate CEFR levels by rounding the fair averages to the nearest whole CEFR level (e.g., 2.25 was rounded down to 2, corresponding A2, and 2.65 up to 3 or B1).

Our decision to combine in the analysis the three writing tasks that each learner wrote, rather than analyse them separately, was based on two related considerations. First, the study contributes to research on the linguistic characteristics of the CEFR *proficiency* levels (e.g., Bartning et al. 2010 and the studies reviewed above). Thus, the focus was on what characterises learners' writing whose writing ability has been assessed to correspond to particular CEFR levels. Second, our perspective is that of language assessment, where it is common to use multiple tasks to increase the reliability and generalizability of the ability estimates. For example, van den Bergh et al. (2012: 23) state that "to measure writing skills reliably, one needs multiple assignments rated by multiple raters". Incidentally, the developers of the TOEFL iBT found that three tasks were required for obtaining adequate reliability (Chapelle 2008: 331).

Rating quality was investigated by examining raters' Infit Mean Square values, which should usually range from 0.6 to 1.5 (e.g., Engelhard 1994). Rater fit was considered to be appropriate as all Infit Mean Square statistics were smaller than 1.3. All point-biserial estimates of the raters were optimistic and between .27 and .65 (for 9 of the 11 raters, they exceeded .42). This suggests that the raters applied the scale in a relatively consistent way, although their severity varied. However, since Facets adjusts the ability measures by taking into account rater severity, these differences did not prevent a reliable estimation of learners' EFL writing ability, mainly when the ability measures were based on three writing tasks. After rating, the handwritten scripts were transcribed for automated analyses.

3.5 Modification of the texts

The scripts were slightly modified for automated analyses. Misspelt words were corrected to allow the tools to identify words correctly, and any missing periods were added to the end of sentences to ensure correct identification of sentence boundaries. Other punctuation, grammatical errors or incorrect word choices were not corrected (on data cleaning, see McNamara et al. 2014: 155-6). No texts were removed from the corpus in the rating and data cleaning stages.

3.6 Linguistic analysis of learners' writing

Each script was investigated with two tools designed to analyse English texts: the L2 Syntactic Complexity Analyzer and Coh-Metrix. *L2 Syntactic Complexity Analyzer* (L2SCA) is a freely available UNIX-based research tool that calculates 14 SC indices (see table 1 and Lu 2010). L2SCA consists of three components: a parser (Stanford parser), a procedure for counting the production units, and an SC analyser. From many *Coh-Metrix* indices, we chose 16 that relate to SC (see table 2 and Graesser et al. 2004).

Table 1. Syntactic complexity indices in the L2 Syntactic Complexity Analyzer based on Lu (2010).

Syntactic complexity	**Index**	**Definition**
Length of production units	Sentence length (mean & standard deviation)	the number of (#) words/# sentences.
	T-unit length	# words/# T-units.
	Clause length	# words/# clauses
	Clauses per sentence	# clauses/# sentences
Sentence complexity	T-unit complexity ratio	# clauses/# T-units
Subordination	Complex T-unit ratio	# complex T-units/# T-units
	Dependent clause ratio	# dependent clauses/# clauses
	Dependent clauses per T-unit	# dependent clauses/# T-units
	Coordinate phrases per clause	# coordinate phrases # clauses
Coordination	Coordinate phrases per T-unit	# coordinate phrases/# T-units
	Sentence coordination ratio	# T-units/# sentences
	Complex nominals per clause	# complex nominals/# clauses
Particular structures	Complex nominals per T-unit	# complex nominals/# T-units
	Verb phrases per T-unit	# verb phrases/# T-units

Table 2. Syntactic complexity indices in Coh-Metrix based on Graesser et al. (2004)

Syntactic complexity Indices	Definition of indices
Syntactic simplicity (z-score & percentile)	The degree to which sentences contain fewer vs more words and use simple vs complex syntactic structures.
Left embeddedness	Mean the number of words before the main verb. These are often structurally dense, syntactically ambiguous, or ungrammatical (Graesser et al. 2004) and difficult to process.
Modifiers per noun phrase	Mean # modifiers/noun phrases.
Minimal edit distance for parts of speech	Combination of semantic and syntactic dissimilarity and distance between parts of speech across sentences (McCarthy et al., 2009).
Sentence syntax similarity (adjacent sentences)	Degree of uniformity and consistency of the syntactic constructions.
Syntactic pattern density indices	Density index (e.g., Noun density, Verb density, Adverbial density, and Preposition phrase density; Negation density, Gerund density, or Infinitive density) / per 1000 words.

3.7 Statistical analyses

Pearson correlation coefficients were used to investigate the relationship between SC indices and writing proficiency ratings (i.e., learner ability measures from Facets). To examine the differences between learners placed at different CEFR levels, several MANOVAs were run on groups of independent variables (i.e., count variables, SC variables from L2SCA and Coh-Metrix) to investigate overall differences between CEFR levels. These were followed by univariate tests (in MANOVA) to examine differences between adjacent CEFR levels. Bonferroni correction was applied to control for the familywise error rate associated with the pairwise comparison of several groups (CEFR levels).

4 Results

Table 3 the distribution of the learners' overall writing ability across CEFR levels, based on rounding Facets fair average values to the nearest whole CEFR level.

Table 3. Learners' EFL writing ability in the two student groups as CEFR levels

Student group / CEFR level	A1	A2	B1	B2
Grade 8	37	87	70	8
Gymnasium	-	31	125	39

Note: One gymnasium student whose fair average was close to A1 is included in A2; two gymnasium students who are close to C1 are included in B2.

The ability to write in English varied considerably among the eighth graders despite having studied the language at school since grade three. The most significant proportion (43%) were at A2 and many also at B1 (35%), but quite a few were still at A1 (18%), and only some at B2. In contrast, almost two thirds (64%) of the gymnasium students were at B1, and the rest at A2 or B2 (16% and 20%, respectively). The higher and more

homogeneous results achieved by the gymnasium students is explained by the fact that they had studied English three years longer and that gymnasia are attended mainly by the more academically oriented students.

4.1 Relationship between syntactic complexity and writing ability

To address Research Question 1 (is SC and writing ability related in Finnish EFL learners), correlation coefficients were computed between the SC indices obtained from the two computer tools and the writing ability measures from Facets. First, we report the correlations between the number of different kinds of linguistic units in learners' writing and their writing ability (see Table 4; Figures 1 and 2). The number (count) of such units – words, clauses, T-units, sentences – indicates text length, which has been found to relate to ratings of L2 writing quality: longer texts are generally considered better than shorter texts and are awarded higher ability ratings. The specific reason for investigating this here was to see if the correlations in both age groups were equally strong.

Table 4. Correlations between count variables and EFL writing ability

Index (Number of …)	Grade 8	Gymnasium
Words	.822***	.621***
Sentences	.573***	.247***
Clauses	.726***	.472***
T-units	.622***	.335***
Dependent clauses	.633***	.283***
Coordinate phrases	.283***	.317***
Complex T-units	.612***	.318***
Complex nominals	.625***	.594***

*** p<.001

The most detailed index of text length, the number of words, correlated strongest with writing ability in both groups (.822 in grade 8; .621 in the gymnasium). However, counts of all other linguistic units also correlated significantly (at p <.001 level) with ability in both groups. Another strong correlation was the number of complex nominals (.625 and .594 in grade 8 and gymnasium, respectively) and clauses (.726 and .472, respectively). There were also differences across the groups: the largest was the sentence count (.573 in grade 8 but only .247 in gymnasium) and the number of dependent clauses (.633 vs .283, respectively). However, the most notable difference was that the correlations across all count variables were significantly more significant in grade 8 (the only exceptions were coordinate phrases and complex nominals). The amount of language produced by the learners, irrespective of the unit of analysis, was a more significant correlation of writing ability in grade 8, whereas its importance was more negligible in the more able gymnasium group.

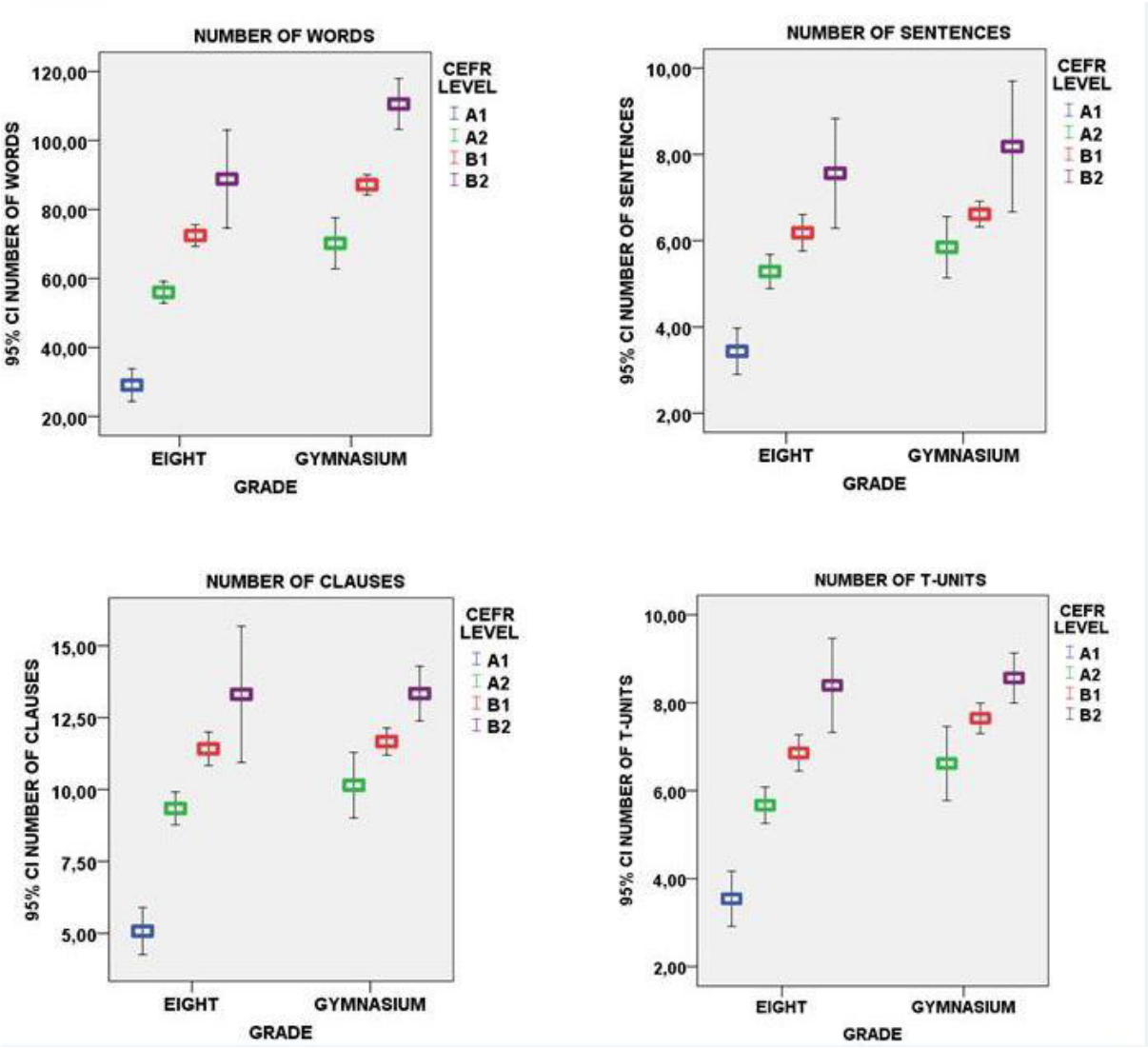


Figure 1. Error-bar charts for the essential count variables at different CEFR levels

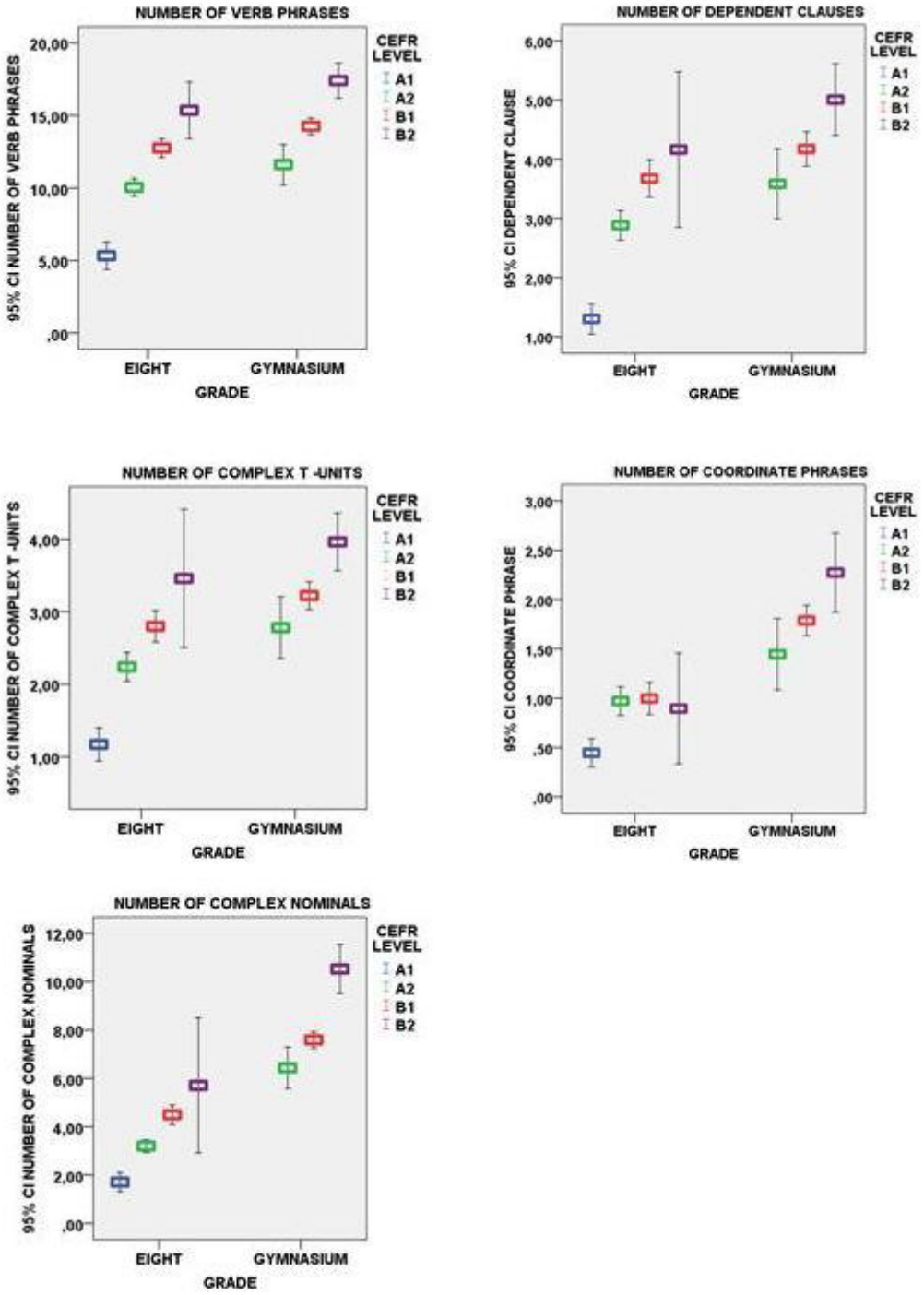


Figure 2. Error-bar charts for the more complex count variables at different CEFR levels

However, to more comprehensively address RQ1, we investigated indices representing different aspects of SC (see Table 1 & 2). The indices in table 5 concern the length of production units. They are typically operationalised as mean lengths of clauses, T-units and sentences, and as their standard deviations.

Table 5. Correlations between mean length of production units and EFL writing ability

Index	Grade 8	Gymnasium
Sentence length (mean)	.429***	.238**
Sentence length (st.dev.)	.495***	.216**
T-unit length (mean)	.321***	.195**
Clause length (mean)	.260***	.375***

** .01 ≥ p ≥ .001 / *** p<.001

Table 5 reports the correlations between measures relating to the length of production units and writing ability measures from Facets. All correlations are statistically significant but low or moderate. Again, most correlations are more robust for the 8th graders, particularly those concerning sentence length (over .4 for both the mean length and variability in sentence length) and T-unit length. However, the mean length of clauses was a more robust correlate of writing ability in gymnasium than in grade 8.

Table 6. Correlations between measures of subordination and coordination and EFL writing ability

Type of index	Index	Grade 8	Gymnasium
Sentence complexity	Clauses per sentence	.220*	ns
Subordination	Clauses per T-unit	.174*	ns
	Complex T-units per T-unit	ns	ns
	Dependent clauses per clause	.224**	ns
	Dependent clauses per T-unit	.140*	ns
Coordination	Coordinate phrases per clause	ns	ns
	Coordinate phrases per T-unit	ns	ns
	T-units per sentence	.187**	ns
Particular structures	Complex nominals per clause	ns	.286***
	Complex nominals per T-unit	ns	.198**
	Verb phrases per T-unit	.262***	ns

* .05 ≥ p ≥ .01 / ** .01 ≥ p ≥ .001 / *** p<.001

The measures of subordination and coordination differed between the groups (see Table 6). Almost all subordination measures correlated modestly with writing ability in grade 8, but no correlations were found for gymnasium. The highest correlations in grade 8 were found for verb phrases per T-unit (.262), dependent clauses per clause (.224), and clauses per sentence (.220). Of coordination measures, only the ratio of T-units per sentence had a small significant correlation with writing ability in grade 8. Particular SC structures were also related to the ratings of writing ability: the number of complex nominals per clause and per T-unit in the gymnasium and verb phrases per T-unit in grade 8.

Table 7. Correlations between measures of syntactic similarity and simplicity and EFL writing ability

Index	Grade 8	Gymnasium
Syntactic simplicity (z-score)	ns	-.206**
Syntactic simplicity (percentile)	ns	-.233**
Left embeddedness	.188**	.247**
Modifiers per noun phrase	ns	.433***
Sentence syntax similarity (adjacent sentences)	-.214**	-.282***

** .01 ≥ p ≥ .001 / *** p<.001

Some Coh-Metrix indices capture variation in the syntactic simplicity and similarity (within paragraphs) and the number of modifiers per the main word in sentences. The findings indicate that these indices relate more strongly to writing ability in the more able gymnasium group, where all indices correlated significantly. Modifiers per noun phrase had the highest correlation (.433), but, interestingly, no significant correlation was found for grade 8. Syntactic simplicity and similarity indices all correlated over .2 with writing ability in the gymnasium, as did left embeddedness. Only the syntactic similarity measures and left embeddedness correlated with writing in grade 8, but only modestly (around .2 or lower). The negative correlations in Table 7 indicate that syntactically similar and straightforward (i.e., lacking variation across the text) was associated with lower writing ability.

4.2 Syntactic complexity as a way to distinguish CEFR writing ability levels

To address Research Question 2 on whether certain syntactic complexity features distinguish specific CEFR levels, multivariate analyses of variance were used to compare SC features across the levels. Tables A and B in Appendix 1 present the means and standard deviations for the count variables for the two learner groups. The counts were calculated with the L2SCA. As the relatively high correlations between count variables and writing ability suggested, the number of words, clauses, sentences and phrases increased steadily across levels (see Figure 1 & 2). Tables 8 and 9 summarise the results of multivariate analyses of variance with the CEFR writing level as the independent variable and the counts of various linguistic units as dependent variables. It should be noted that an omnibus Manova analysis was first conducted to the indices reported in each table; in each case, the results were statistically significant, which then warranted the univariate analyses of each SC index reported as the overall F- and p-values, as well as effect sizes in Tables 8–13.

Table 8. Count variables: summary of statistical significance of overall and between CEFR level differences in grade 8

Index (Number of …)	**A1 vs A2**	**A2 vs B1**	**B1 vs B2**	**Overall**		
	p.	*p.*	*p.*	*F*	*p.*	η^2
Words	<.001	<.001	.014	84.789	<.001	.57
Sentences	<.001	.011	.206	27.285	<.001	.26
Clauses	<.001	<.001	.262	56.479	<.001	.46
T-units	<.001	<.001	.145	35.675	<.001	.35
Verb Phrase	<.001	<.001	.066	66.52	<.001	.50
Dependent clauses	<.001	.001	.830	51.994	<.001	.45
Complex T-units	<.001	.001	.255	38.909	<.001	.37
Coordinate phrases	<.001	.995	.976	11.624	<.001	.15
Complex nominals	<.001	<.001	.747	34.257	<.001	.34

Note: After Bonferroni correction for 4 CEFR levels, only those pairwise comparisons where p≤.008 can be considered significant

Table 9. Count variables: Summary of statistical significance of overall and between CEFR level differences in the gymnasium

Index (Number of …)	**A2 vs B1**	**B1 vs B2**	**Overall**		
	p	p	F	p	η^2
Words	<.001	<.001	42.82	<.001	.31
Sentences	.367	.003	7.96	.016	.08
Clauses	.022	.004	11.45	<.001	.11
T-units	.031	.036	8.36	.001	.08
Verb Phrase	<.001	<.001	25.31	<.001	.21
Dependent clauses	.229	.023	6.45	.005	.06
Complex T-units	.147	.001	10.36	<.001	.10
Coordinate phrases	.193	.068	6.59	.012	.06
Complex nominals	.038	<.001	33.64	<.001	.26

Note: After Bonferroni correction for 3 CEFR levels, only those pairwise comparisons where p≤.017 can be considered significant

Tables 8 and 9 show that, overall, all count indices separated the CEFR levels significantly. Separation was more apparent in grade 8, as indicated by larger effect sizes than in the gymnasium. The tables also display that the number of words learners wrote differed between almost all adjacent CEFR levels. However, almost all count variables are distinguished between A1 and A2 writers on the one hand and between A2 and B1 writers on the other in grade 8. In contrast, these variables did not clearly distinguish A2 and B1 in the gymnasium but did a better job separating B1s from B2s, particularly the number of complex nominals, complex T-units, clauses, and sentences.

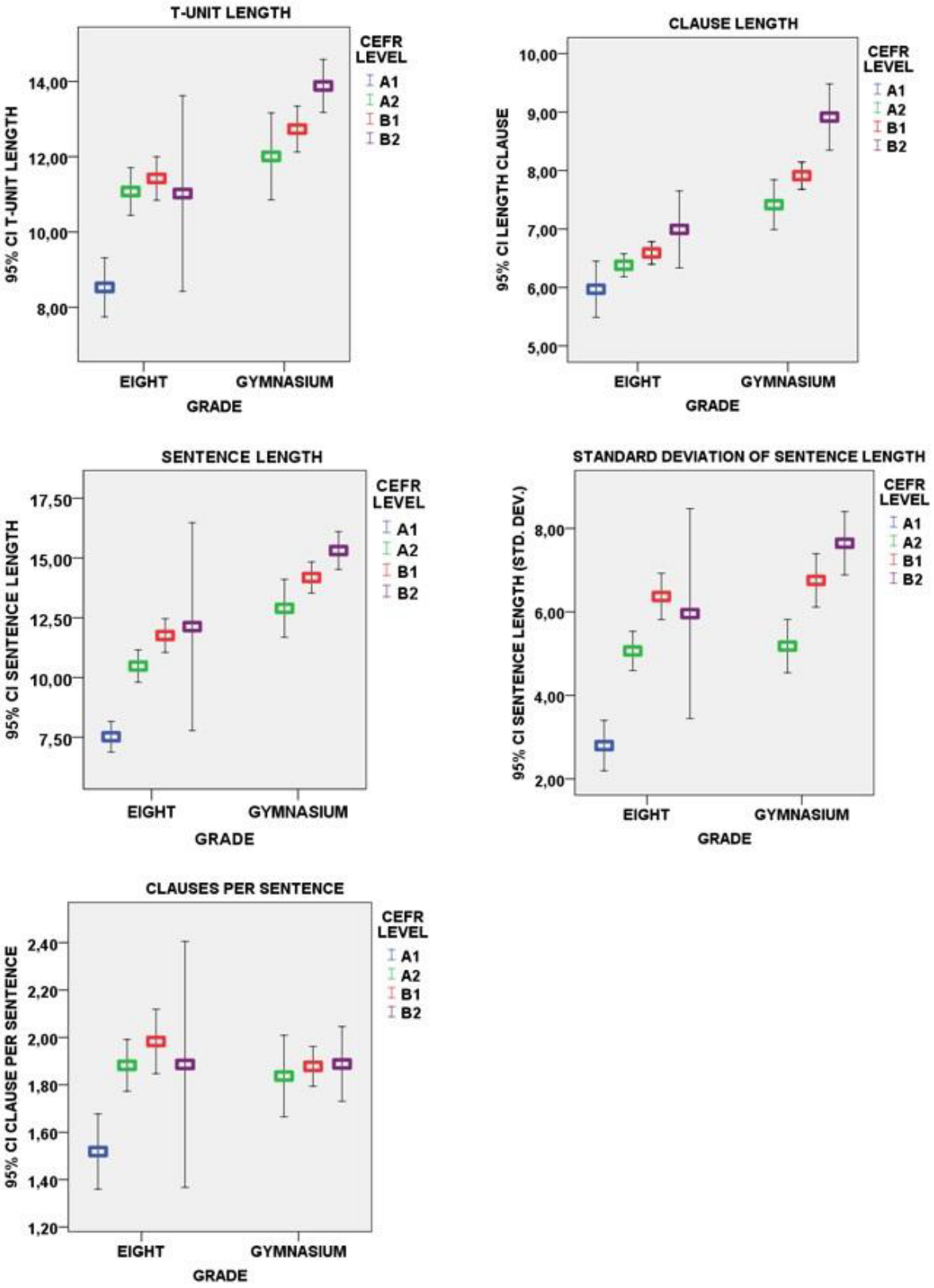


Figure 3: Error-bar charts for the mean clause, T-unit and sentence length, and clauses per sentence at different CEFR levels

Tables C and D (Appendix 1) display the means and standard deviations across the CEFR levels for the SC variables obtained from L2SCA. Tables show the mean length of the production units increasing from level to level (Figure 3). A similar trend can be seen for sentence complexity (clauses per sentence) and such structures as the number of complex nominals per clause or T-unit.

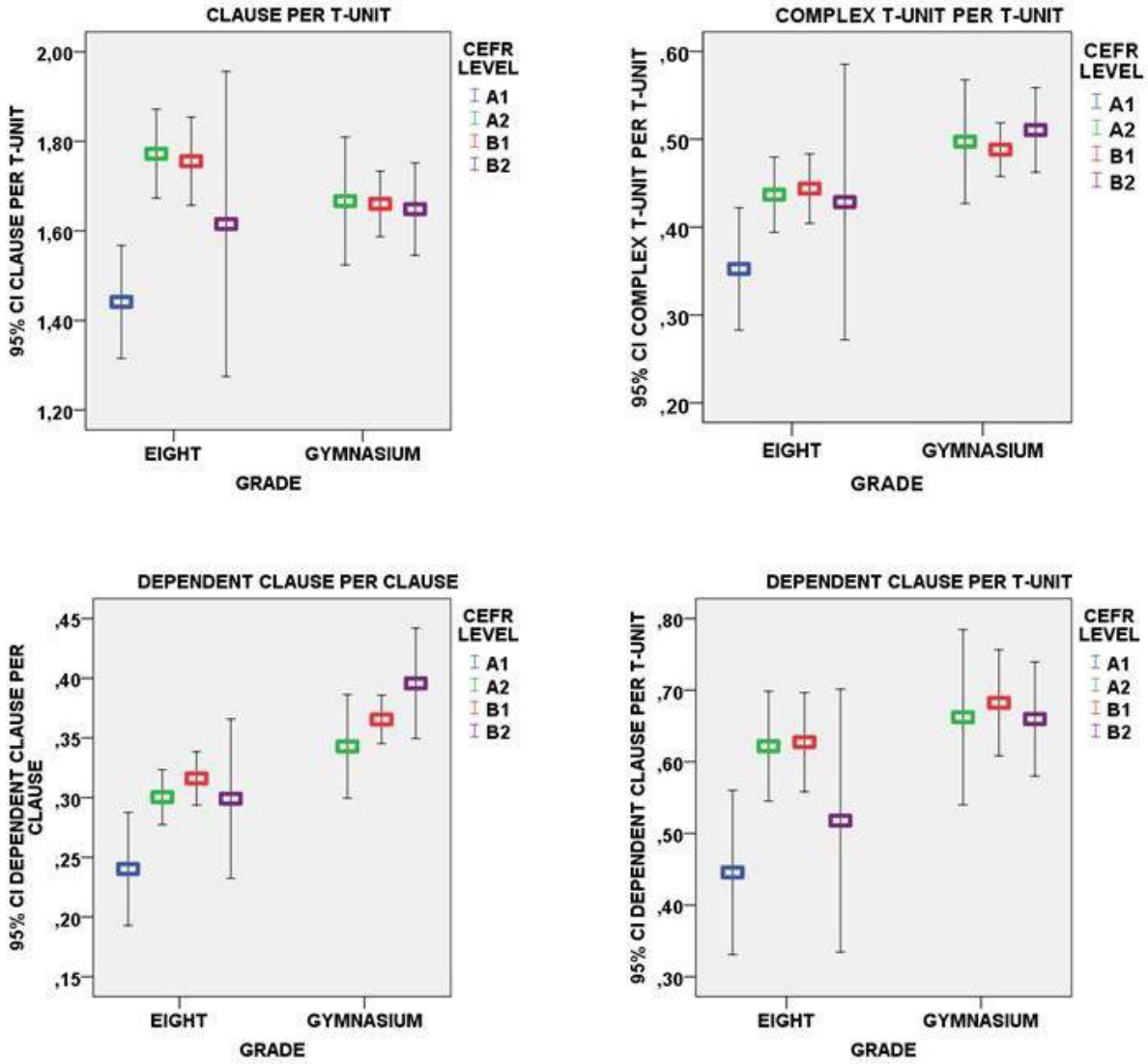


Figure 4: Error-bar charts for subordination indices at different CEFR levels

Tables 10-11 report the statistical significance of the differences for the SC variables obtained from L2SCA both overall (across CEFR levels) and between adjacent CEFR levels (see also Figure 3). The length of the production units separated the levels significantly: Sentence and T-unit length distinguished A1 vs A2 and clause length B1 vs B2. Sentence complexity increased significantly from A1 to A2. Similarly, the only significant subordination index (clauses per T-unit) distinguished between A1 and A2 but not above. Two coordination indices separated CEFR levels overall but failed to distinguish between adjacent levels. In contrast, particular syntactic structures turned out to be significant: the number of verb phrases per T-unit distinguished A1 and A2, whereas the number of complex nominals per clause separated B1 from B2.

Table 10. Syntactic complexity indices from L2SCA: summary of statistical significance of overall and between CEFR level differences in grade 8

Index	**A1 vs A2**	**A2 vs B1**	**B1 vs B2**	**Overall**		
	p.	*p.*	*p.*	*F*	*p.*	η^2
Sentence length	<.001	.081	1.000	24.540	.000	.20
T-unit length	<.001	.962	.999	12.491	.000	.14
Clause length	.392	.428	.548	3.306	.033	.06
Clauses per sentence	.004	.803	.997	6.982	.001	.09
Clauses per T-unit	.001	1.000	.944	6.618	.001	.08
Complex T-units per T-unit	.140	1.000	1.000	1.806	.167	.03
Dependent clauses per clause	.034	.938	.999	2.751	.059	.06
Dependent clauses per T-unit	.041	1.000	.939	2.955	.047	.05
Coordinate phrases per clause	.574	.312	.695	1.775	.171	.02
Coordinate phrases per T-unit	.031	.602	.858	3.971	.015	.03
T-units per sentence	.494	.083	.998	3.628	.023	.06
Complex nominals per clause	1.000	.959	1.000	.271	.846	.004
Complex nominals per T-unit	.349	.907	.998	2.046	.128	.03
Verb phrases per T-unit	<.001	.991	.998	8.308	.000	.10

Note: After Bonferroni correction, only those pairwise comparisons where p≤.008 can be considered significant

Table 11. Syntactic complexity indices from L2SCA: summary of statistical significance of between and overall CEFR level differences in the gymnasium

Index	**A2 vs B1**	**B1 vs B2**	**Overall**		
	p.	*p.*	*F*	*p.*	η^2
Sentence length	.265	.183	5.50	.006	.043
T-unit length	.590	.146	5. 02	.009	.032
Clause length	.106	.005	11.36	<.001	.110
Clauses per sentence	.964	.999	.114	.893	.001
Clauses per T-unit	1.000	.998	.020	.974	.0002
Complex T-units per T-unit	.991	.857	.259	.772	.002
Dependent clauses per clause	.728	.440	1.42	.183	.017
Dependent clauses per T-unit	.991	.983	.073	.929	.0007
Coordinate phrases per clause	.942	.810	.560	.572	.005
Coordinate phrases per T-unit	.980	.877	.353	.703	.003
T-units per sentence	.612	.738	.718	.489	.007
Complex nominals per clause	1.000	.001	6.906	.001	.070
Complex nominals per T-unit	1.000	.042	4.070	.022	.031
Verb phrases per T-unit	.891	.720	1.369	.261	008

Note: After Bonferroni correction, only those pairwise comparisons where p≤.017 can be considered significant

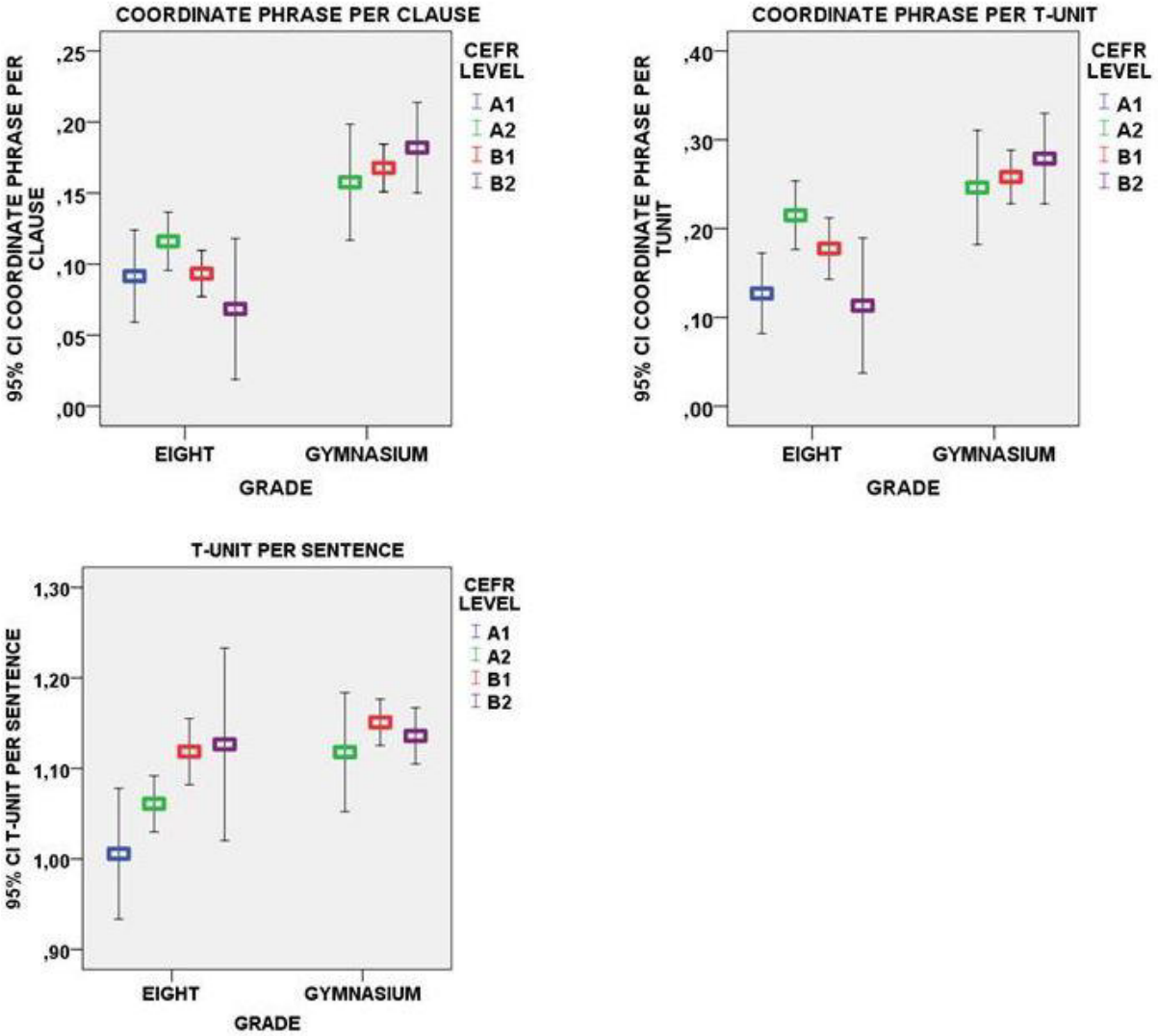


Figure 5: Error-bar charts for coordination indices at different CEFR levels

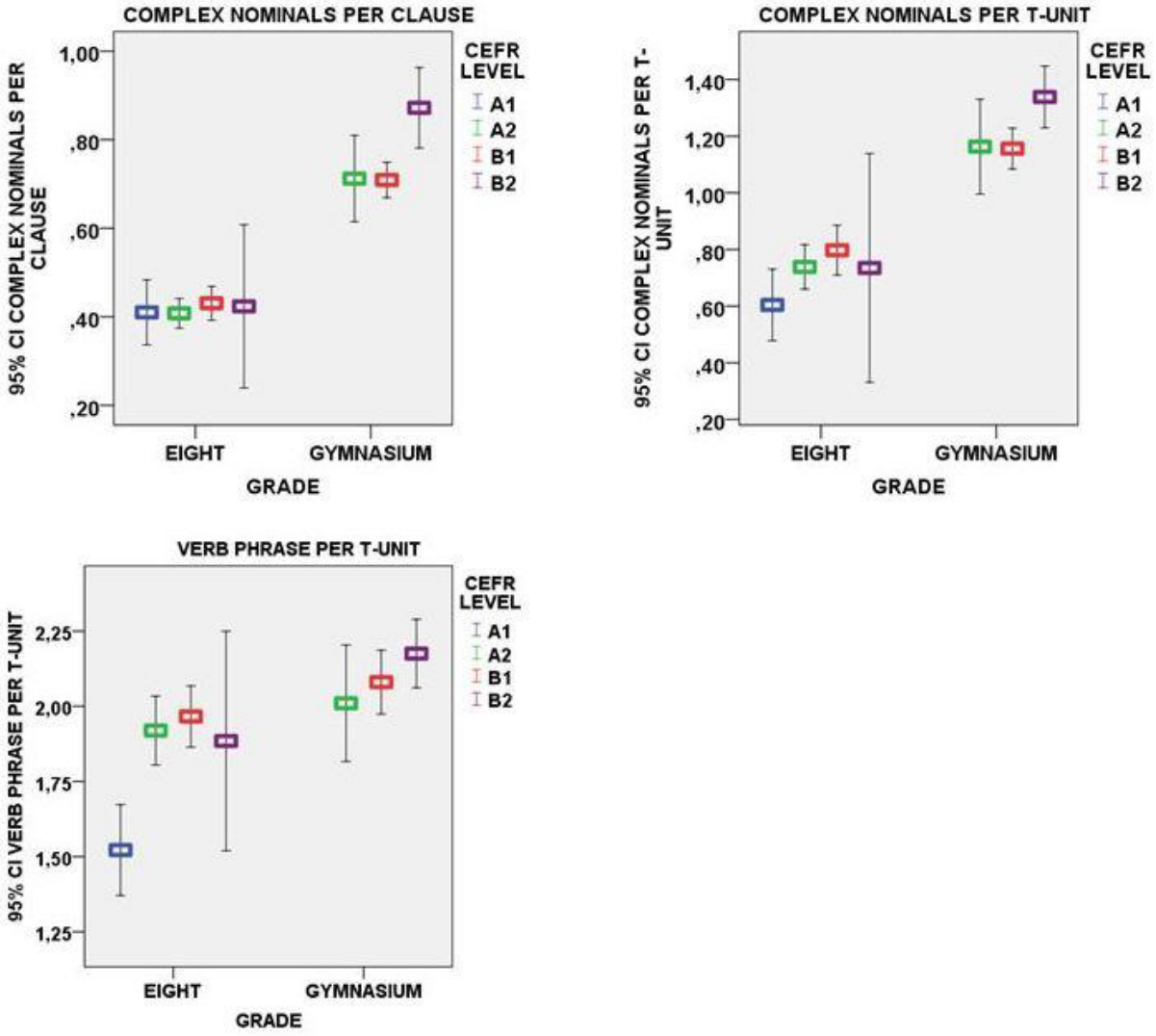


Figure 6: Error-bar charts for particular structures at different CEFR levels

Finally, we report on the results for the somewhat different SC indices from Coh-Metrix (see table E and F in Appendix 1 for the means and standard deviations). Coh-Metrix reports both the mean sentence length and its standard deviation. The tables show that the mean standard deviation of average sentence length primarily increased from level to level. Syntactic simplicity indices had a slight downward trend implying that syntax becomes more complex as proficiency improves. A similar trend can be seen for syntactic similarity. Left embeddedness and the number of modifiers per noun phrase increased slightly from level to level. Density measures displayed both downward (noun and negative phrase density) upward trends (adverbial, preposition and passive voice density).

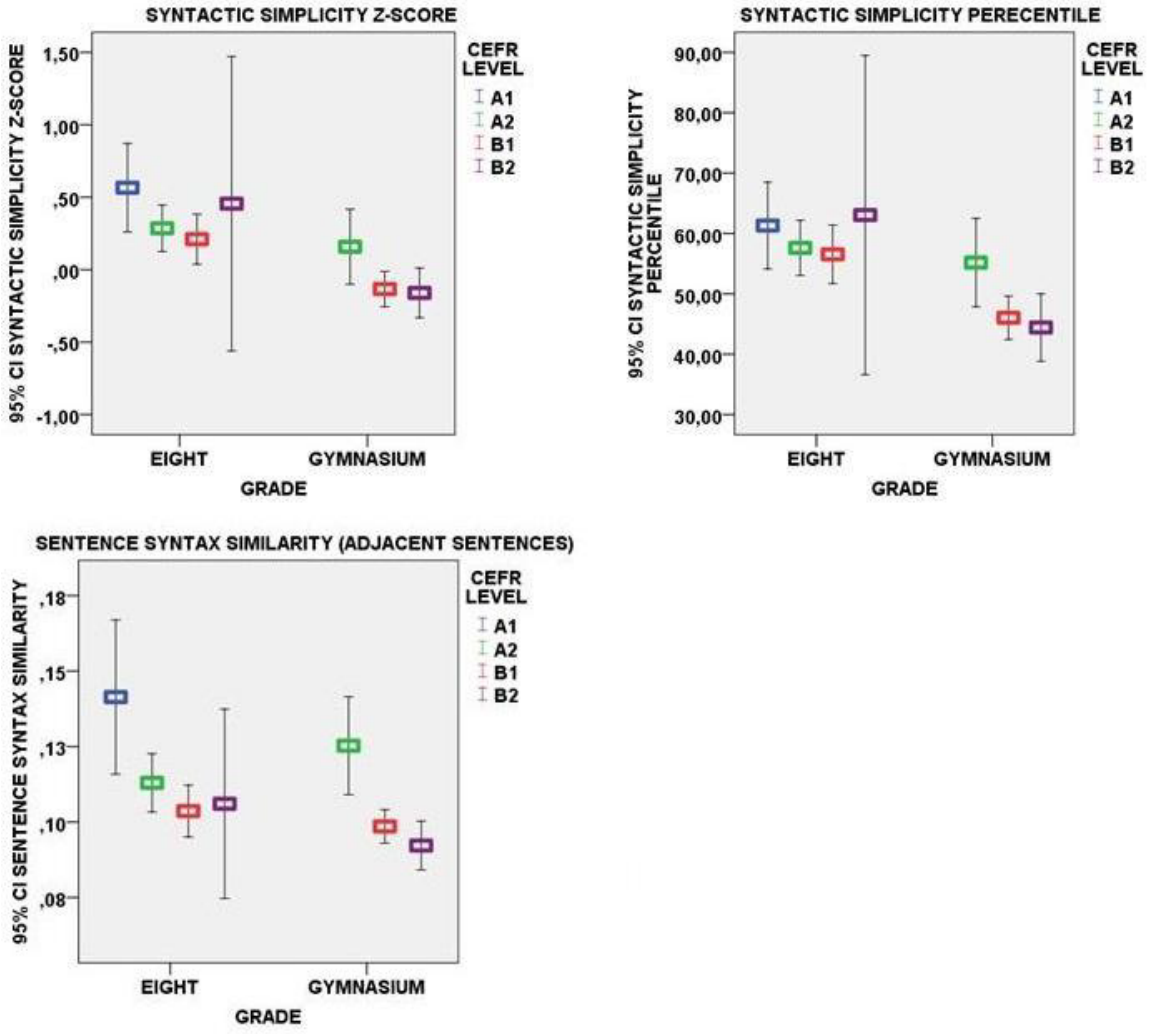


Figure 7: Error-bar charts for syntactic simplicity and sentence syntax similarity indices at different CEFR levels

Tables 12-13 report the statistical significance of the overall and between-level differences in the SC variables from Coh-Metrix (see also Figures 8, 9 & 10). The standard deviation of the mean sentence length separates the three lowest levels (A1-B1) mainly. Overall syntactic simplicity decreased from lower to higher levels (particularly in the gymnasium), but none of the adjacent levels was separable. Sentence syntax similarity indices distinguished CEFR levels more clearly, but the only significant pairwise difference was found between A2 and B1 (in the gymnasium). Left embeddedness and the number of modifiers per noun phrase separated B1 and B2 levels but not below. The minimal edit distance for parts of speech separated A1 and A2 but not beyond. Of the density measures, only infinitive and noun phrase densities distinguished CEFR levels; the former between A1 and A2, between A2 and B1, and the latter between A2 and B1, all in grade 8.

Table 12. Syntactic complexity indices from Coh-Metrix: summary of statistical significance of overall and between CEFR levels differences in grade 8

Index	**A1 vs A2**	**A2 vs B1**	**B1 vs B2**	**Overall**		
	p.	*p.*	*p.*	*F*	*p.*	η^2
Sentence length (st.dev.)	<.001	.002	.997	21.337	<.001	.24
Syntactic simplicity (z-score)	.372	.992	.958	1.739	.160	.03
Syntactic simplicity (percentile)	.943	1.000	.961	.552	.647	.008
Left embeddedness	.026	1.000	1.000	3.434	.018	.05
Modifiers per noun phrase	.418	.930	1.000	1.020	.385	.02
Sentence syntax similarity (adjacent sentences)	.166	.474	.998	18.412	<.001	.22
Minimal edit distance for PoS	.004	<.001	.999	4.152	.007	.05
Noun phrase density	.844	.001	.985	5.268	.002	.07
Verb phrase density	.045	.418	.990	6.868	<.001	.09
Adverbial phrase density	.765	.177	.875	2.942	.034	.04
Preposition phrase density	.120	1.000	.668	3.621	.014	.05
Negation density	.300	1.000	.971	1.848	.140	.03
Gerund density	.929	.024	.603	2.854	.038	.04
Infinitive density	<.001	<.001	.826	27.128	<.001	.29

Note: After Bonferroni correction, only those pairwise comparisons where p≤.008 can be considered significant

Table 13. Syntactic complexity indices from Coh-Metrix: summary of statistical significance of overall and between CEFR level differences in the gymnasium

Index	**A2 vs B1**	**B1 vs B2**	**Overall**		
	p.	*p.*	*F*	*p.*	η^2
Sentence length (st.dev.)	.041	.331	13.63	.006	.05
Syntactic simplicity (z-score)	.085	.995	2.66	.072	.03
Syntactic simplicity (percentile)	.064	.958	3.15	.045	.03
Left embeddedness	.948	.001	8.10	<.001	.07
Modifiers per noun phrase	.090	.000	14.26	<.001	.13
Sentence syntax similarity (adjacent sentences)	.008	.397	10.33	<.001	.10
Minimal edit distance for PoS	.908	.035	3.80	<.027	.02
Noun phrase density	.046	.678	4.22	.016	.04
Verb phrase density	.989	.999	0.06	.939	.006
Adverbial phrase density	.045	.956	3.82	.024	.04
Preposition phrase density	.240	.195	4.49	.012	.04
Negation density	.281	.704	2.25	.108	.02
Gerund density	.511	.112	3.70	.027	.03
Infinitive density	.915	.245	1.90	.148	.10

Note: After Bonferroni correction, only those pairwise comparisons where p≤.017 can be considered significant

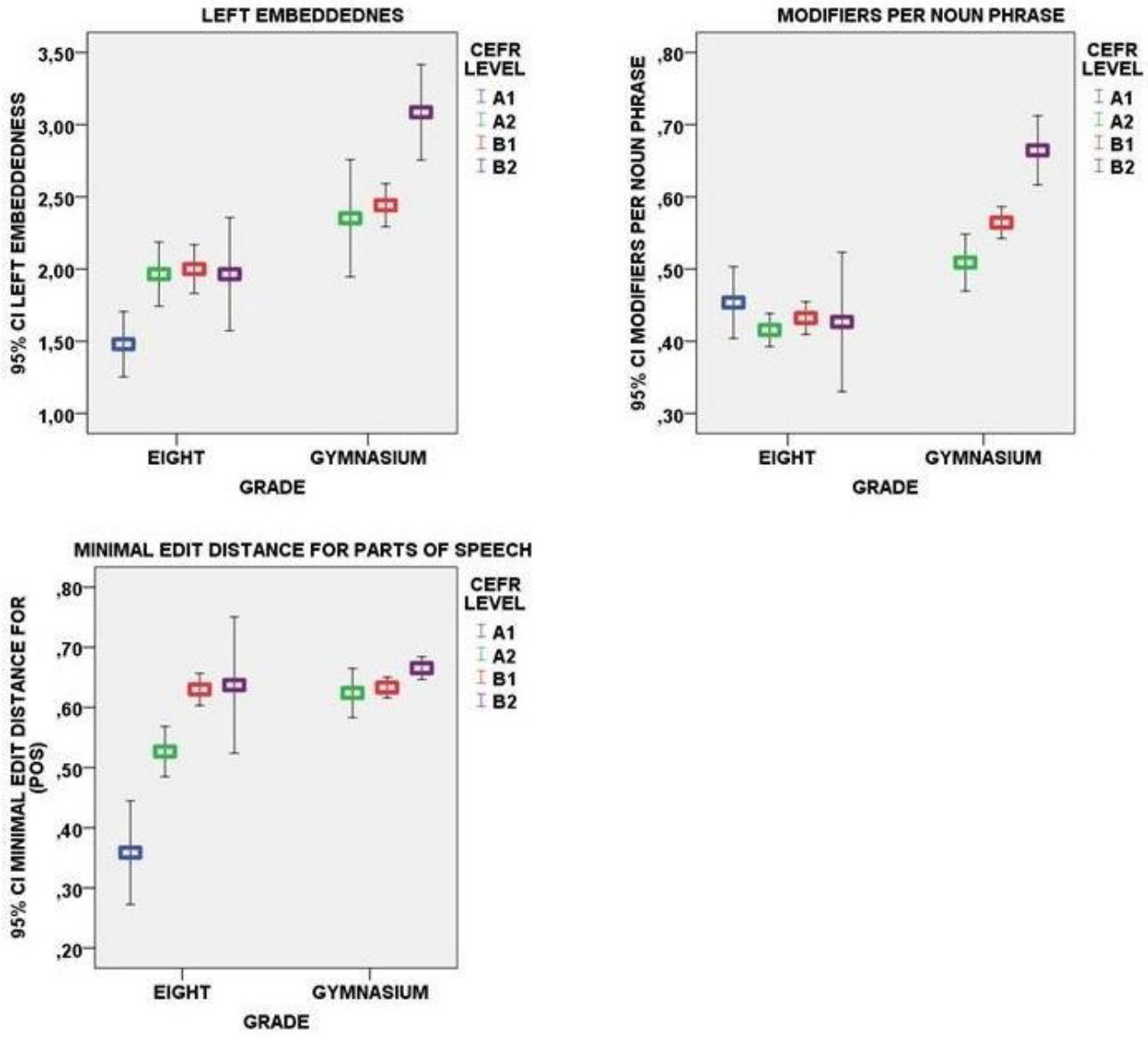


Figure 8: Error-bar charts for syntactic complexity indices from Coh-Metrix at different CEFR levels

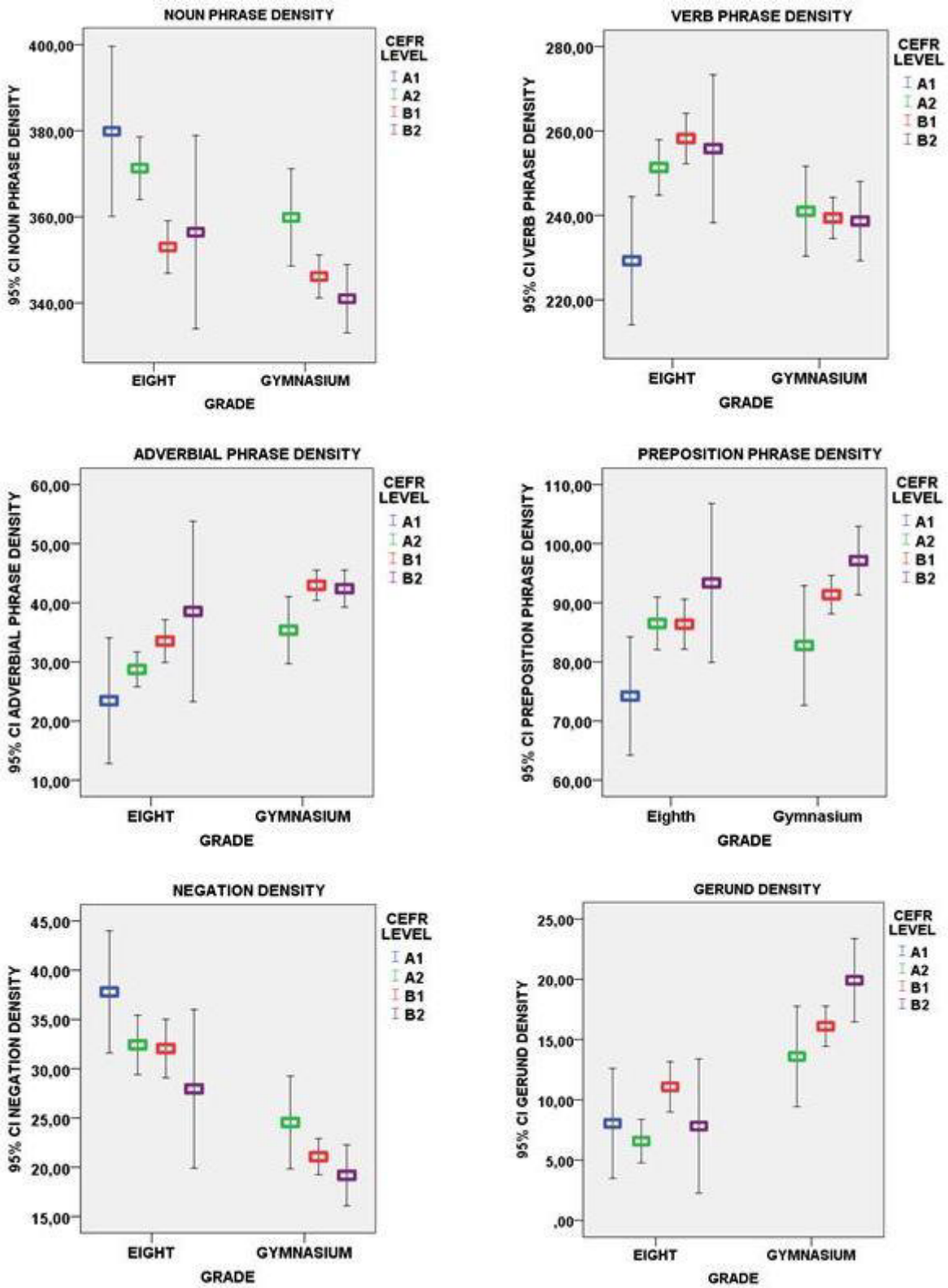


Figure 9: Error-bar charts for syntactic pattern density indices at different CEFR levels

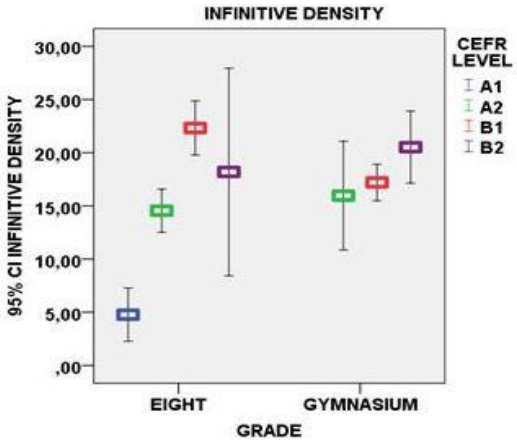


Figure 10: Error-bar chart for Infinitive Density (Syntactic pattern density) at different CEFR levels

5 Discussion

The study sheds light on the linguistic characteristics of the CEFR levels by focusing on syntactic complexity in the writing of two groups of Finnish-speaking EFL learners aged 14 and 17. The groups also differed in terms of proficiency: the writing ability of the older gymnasium students was higher since they had studied English longer. Therefore, the comparison of A1 and A2 levels was possible only for the 8th graders as there were no A1 writers in the gymnasium. For its part, the comparison between B1 and B2 was possible, in practice, only among the gymnasium students since there were only eight B2 writers in grade 8.

Our RQ1 concerned the relationship between syntactic complexity in the learners' writing and their writing ability, based on three double-rated writing tasks, and whether the results varied across the two groups. First, we found that text length (number of words, clauses, sentences, etc.) correlated strongly with the ability (even over .8); the correlations were more substantial in the younger group. This suggests that raw text length may be a more vital indicator of L2 writing ability in the early stages of L2 learning, but then its role diminishes but may not disappear, not at least before B2. As for the actual indices of SC, we found that the length of production units (e.g., clauses, sentences) correlated significantly but only moderately with writing quality and more strongly among the 8th graders. The findings confirm the expectation that simple counts of linguistic units are often quite good predictors of learners' L2 (writing) ability, including counts of such indices of SC as dependent clauses and complex nominals and complex T-units, even if there appear to be differences that relate to learners' age and/or ability.

We discuss the second RQ (whether SC separates CEFR levels) below and compare the findings with previous research. There are still relatively few studies on the relationship between SC in EFL writing and CEFR levels. Table 14 summarises the significant differences in SC between CEFR levels in both our study and previous research. A direct comparison of our findings with those reported previously is complicated since the SC indices investigated and how the results are reported may vary.

Such caveats notwithstanding, Table 14 allows us to compare different studies and examine trends in research on SC. The present study is referred to with the letter 'A' in Table 14, and A8 refers to grade 8 and AG to the gymnasium. The previous studies are numbered from one to nine (see the key after the table).

Table 14. Summary of significant differences in syntactic complexity across CEFR levels in the current and previous studies

Syntactic complexity indices used in this study	**CEFR levels separated in a particular study**				
	A1 / A2	**A2 / B1**	**B1 / B2**	**B2 / C1**	**C1/C2**
Sentence length	A8, 1, 2	1, 2, 5, 7, 9	1, 5, 9	1, 5	1, 5
Sentence length (st. dev.)	A8, 2	A8, (2)			
T-unit length	A8, (2), (8)	2, 4, 8, 9	9		
Clause length		1, 4, 9	1,9,AG		
Clauses per sentence	A8, 2	(9)			
Clauses per T-unit	A8, (2)	1, 4, 9	6		
Complex T-units /T-unit	(2)	(2), 9	(9)		
Dependent clauses / clause	(2)	2, (9)	6, (9)		
Dependent clauses /T-unit	1, (2)	1, (9)	1,6, (9)		
Coordinate phrases /clause					
Coordinate phrases /T-unit					
T-units /sentence	(2)				
Complex nominals /clause		9	9, AG		
Complex nominals /T-unit		(2),9	9		
Verb phrases /T-unit	A8, 2	2, (9)	9		
Syntactic simplicity	(8)				
Left embeddedness	(2)		AG		
Modifiers per noun phrase	(2)	(2)	AG	3	
Minimal edit distance	A8, (2)	A8			
Sentence syntax similarity		AG			3
Noun phrase density		A8, (2)		3	
Verb phrase density					
Adverbial phrase density		AG			
Preposition phrase density	(2)		6		
Negation density	(2)	(2)			
Gerund density	(2)	(2)	6		
Infinitive density	A8, 2	A8, AG, (2)	6		
SC indices not used in the current study:					
The proportion of simple vs complex sentences	8				
Compound and complex sentence ratios		7			
Coordinate and dependent clause ratios		7			
Number of finite relative clauses		8			
Adverbial, adjective & nominal clauses per clause			6		
Noun phrases per clause		7			

Key to the letters and numbers in Table 14:
A. Authors (current study); A8 = 8th grade; AG = gymnasium
1. Alexopoulou et al. 2017
2. Authors (XXXX); 2 in brackets = finding concerns only one of the two L1 groups

3. Green 2012
4. Gyllstad et al. 2014
5. Hawkins & Filipović 2012
6. Kim 2004
7. Martínez 2018
8. Verspoor et al. 2012
9. Barrot & Agdeppa 2021

Overall, Table 14 shows that a wide range of SC indices has been found to distinguish CEFR levels. Mean sentence length is a consistent separator across the entire scale (Alexopoulou et al. 2017, Hawkins and Filipović 2012, Barrot and Agdeppa, 2021). In our study, it was a significant separator of the levels in the overall analysis for both age groups, but only the A1 vs A2 pairwise comparison in grade 8 was significant. However, variation in sentence length (i.e., standard deviation) increased significantly across A1–B1 for grade 8.

T-unit length is a reasonably good separator in the A1–B1 range, whereas clause length seems to distinguish at A2 to B2. The current study partly concurs with these results even though the T-unit length only separated A1 from A2 (grade 8).

Sentence level complexity (clauses or T-units per sentence) has separated only between the two lowest CEFR levels in previous research (partly in this study, too) but other sentence-level indices designed by Bulté and Housen (2014) and employed by Martínez (2018) – that is, compound and complex sentence ratios – distinguished A2 from B1. In addition, Verspoor et al. (2012) reported that the proportion of complex and straightforward sentences separated A1 and A2.

Coh-Metrix includes general indices of syntactic simplicity, similarity and variability, but these appear not to have been investigated widely. Interestingly, Green (2012) found a syntactic similarity to distinguish C1 and C2. We found the same for A2 vs B1 but only in the gymnasium. Furthermore, the Authors (XXXX) found a tendency for syntactic similarity to decrease from A1 to B1, but the adjacent levels could not be significantly separated. In the present study, we found minimal edit distance to distinguish A1 vs A2 vs B1 in grade 8.

A wide range of clause level SC indices has been used previously. Clauses or dependent clauses per T-unit appear to distinguish in the A1–B2 range relatively consistently, but only clauses per T-unit separated only A1 vs A2 in our study. Dependent clauses per clause have also separated across A1–B2 in some previous research, but our study failed to replicate that. Martínez (2018), who used different SC indices from us, found both coordinate and dependent clause ratios and noun phrases per clause to differentiate A2 and B1.

Several indices that are at the phrasal in nature (or perhaps borderline between phrasal and clausal) are included in Coh-Metrix, but apart from the current authors and Barrot and Agdeppa (2021), they have not been widely used in CEFR-related SC research; Barrot and Agdeppa found complex nominals per clause or T-unit to distinguish A2 vs B1 vs B2; we only found complex nominals per clause to separate B1 from B2. One of the most interesting of these is the number of modifiers per noun phrase, which the Authors (XXXX) discovered to be the only SC index to show non-linear development from A1 to B1. It first decreased between A1 and A2 but then increased. In the current study, a comparison of A1 and A2 is only possible in the younger age group where the value for this index indeed decreased from A1 to A2, but the difference was not significant. The older age group increased steadily from A2 and was particularly pronounced between B1 and B2. Taken together, the two studies suggest that even if the

number of modifiers might first decrease, it appears to increase after A2. Green's (2012) finding that this index separates C1 from B2 suggests that the trend continues even beyond B2.

Previous studies on the other phrasal level indices have discovered some of them to separate some CEFR levels. Infinitive density, in particular, seems to distinguish in the A1-B2 range, including our study. Of the other such indices, only left embeddedness distinguished only B1 vs B2 and adverbial phrase density A2 vs B1.

In summary, our study sheds light on which SC indices distinguish the CEFR levels A1–B2, and we can compare these with the results of previous research. The effect sizes (tables 10-13) indicate that the most important indices that separate CEFR levels A1–B2 among the younger, less proficient learners were infinitive density, mean sentence length (and its standard deviation), T-unit length, and sentence syntax similarity across adjacent sentences. For the older, more proficient group, the key indices were the number of modifiers per noun phrase, mean clause length, sentence syntax similarity, edit distance and left embeddedness. Combining these findings with those found in previous research, we can tentatively conclude that the length of the more extended production units (sentences and clauses) and variation in their length are among the critical SC features that separate EFL writing from A1 to B1. What appears to separate B1 from B2 and above is mainly related to complexity at the clausal and phrasal levels.

5.1 Limitations

Some limitations of the study and issues with the comparability of different studies need mentioning. In the literature review, we noted that differences across studies in the SC indices, tasks, learners' age and L1 background, and the reliability of placing writing samples on the CEFR levels are all challenges to comparisons. Automated analyses can also be unreliable. For example, the Charniak parser (Charniak 2010) underlying Coh-Metrix is reported to achieve 89% accuracy with L1 English texts, and Crossley and McNamara (2014) estimate the accuracy is likely lower for learner writing. Furthermore, the relatively short texts that many learners in our study wrote may not always provide sufficient data for reliable extraction of some SC features.

Our study did not investigate differences in SC between the writing tasks as we aimed to obtain a more generalisable picture of SC by combining the results of several writing tasks, which is a standard practice in language testing. This approach ignores task-related differences in SC due to register variation; however, our tasks represented only two broad registers (argumentation and narration), partly addressing this limitation. One additional avenue for future research is; therefore, studies focusing on particular tasks and/or applying the Multidimensional Model paradigm, which has not yet been used in research on the linguistic basis of CEFR levels (see, e.g., Biber et al. 2020), and which has to potential to provide valuable insights into writing development, for example, for diagnostic assessment purposes (Author2 et al., forthcoming).

The number of learners in some groups in our study was relatively small (e.g., there were only eight 8th graders whose writing was estimated to be at B2). We decided to leave them in the analyses simply to find out if any of the SC indices would manifest such significant differences between the B1 and B2 level learners in that age group that the difference would be significant. One such index was indeed found (word count; Table 8), and also, the number of verb phrases came close to being a significant separator of B1 and B2 learners.

Another issue with our study – and all CEFR-related research – is the CEFR scale itself. The scales are not ideal for rating purposes since it is unclear how accurately they describe stages of L2 development (e.g., Hulstijn 2007) and since they describe proficiency in rather general terms, unlike scales explicitly developed for rating. Part of this issue is the uncertainty of how much attention the raters paid to SC when rating the performances, even if the scale descriptions did not directly refer to SC. It should be noted, however, that the Facets analyses indicated that the raters could systematically use the scale to distinguish learners with different writing ability levels. Furthermore, significant and relatively strong correlations between the learners' writing ability and the other EFL measures taken by the learners in the more extensive study (e.g., vocabulary, reading and dictation tests) of which this research was part gives further credibility to the writing ability ratings.

5.2 Future research

Finally, Table 14 displays a state of the art of research on SC in written L2 English and, thus, provides us with suggestions for further research on SC. First, it shows that most research concerns the lower levels of proficiency, from A1 to B1. Hence, less is known about how SC separates between B2, C1, and C2. Second, the table reflects that most studies have covered only a limited set of SC indices and, therefore, the gaps (empty cells or cells with only one entry) in the table are often simply due to lack of attention to the particular SC index in research. More wide-ranging studies of SC indices are needed.
Furthermore, some of the studies suggest that the L1 background of the language learners may impact SC in their L2 English texts: this is indicated by the different findings by the Authors (XXXX) for the two L1 groups. Similarly, the current study resulted in several differences in SC between the two age groups, even in the A2–B1 range. The fact that only one of the three writing tasks that each learner completed was the same in both groups makes it impossible to disentangle possible age and task effects. Nevertheless, a further conclusion is that both learners' age and writing task(s) are possible sources of variation in syntactic complexity and, therefore, should be examined in more detail in the future. One additional direction for research could also be mentioned, namely comparing the syntactic complexity of EFL learners at different CEFR levels with the SC of the same-aged native English speakers. This would provide an additional perspective to SC in writing among EFL learners.

References

Ai, Haiyang & Xiaofei Lu. 2013. A corpus-based comparison of syntactic complexity in NNS and NS university students writing. In Ana Díaz-Negrillo, Nicolas Ballier & Paul Thompson (eds.), *Automatic treatment and analysis of learner corpus data*, 249–264. Amsterdam: John Benjamins.

Alderson, J. Charles. 2007. The CEFR and the need for more research. *The Modern Language Journal 914*. 659–663.

Alexopoulou, Theodora, Marije Michel, Akira Murakami & Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques, *Language Learning, 67*. 180–208.

Aryadoust, Vahid, Li Ying Ng & Hiroki Sayama. 2020. A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. Language Testing, 38. 6–40.
Authors (XXXX)
Author2 et al. (forthcoming)
Barrot, Jessie & Joan Agdeppa. 2021. Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing 47*. 100510.
Bartning, Inge, Maisa Martin & Ineke Vedder. 2010. *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research.* Eurosla.
Biber, Douglas. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes, 15.* 133–163.
Biber, Douglas, Bethany Gray, Shelley Staples & Jesse Egbert. 2020. Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes 46*. 100869
Bulté, Bram & Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken, & Ineke Vedder (eds.), *Dimensions of L2 performance and proficiency Investigating complexity, accuracy and fluency in SLA,* 21–46. Amsterdam: John Benjamins.
Bulté, Bram & Alex Housen. 2014. Conceptualising and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing 26*. 42–65.
Carlsen, Cecilie. 2012. Proficiency level – a fuzzy variable in computer learner corpora. *Applied Linguistics 33*. 161–183.
Chapelle, Carol 2012. The TOEFL validity argument. In Carol Chapelle, Mary Enright & Joan Jamieson (eds.) *Building a validity argument for the Test of English as a Foreign Language*, 319–352. New York: Routledge.
Charniak, Eugene. 2010. Top-down nearly-context-sensitive parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 674–683. Stroudsburg, PA: Association for Computational Linguistics.
Crossley, Scott & Danielle McNamara. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing 26*. 66–79.
Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: CUP.
Council of Europe. 2004. *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment. Manual*. Strasbourg: Language Policy Division.
Engelhard, George. 1994. Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement 31*. 93–112.
Graesser, Arthur, Danielle McNamara, Max Louwerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behaviour Research Methods, Instruments, & Computers 36*. 193–202.
Green, Anthony. 2012. *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge: CUP.
Gyllstad, Henrik, Jonas Granfeldt, Petra Bernardini & Marie Källkvist. 2014. Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. In Leah Roberts, Ineke

Vedder & Jan Hulstijn (eds.) *EUROSLA Yearbook 14*. 1–30. Amsterdam: John Benjamins.
Hawkins, John & Luna Filipović. 2012. *Criterial features in L2 English: Specifying the reference levels of the Common European Framework.* Cambridge: CUP.
Hulstijn, Jan. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal 91*. 663–667.
Hulstijn, Jan, J. Charles Alderson & Rob Schoonen. 2010. Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them. In Ineke Bartning, Maisa Martin & Ineke Vedder (eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research,* 11–20. EuroSLA.
Hawkins, John & Luna Filipović. 2012. *Criterial features in L2 English: Specifying the reference levels of the Common European Framework.* Cambridge: CUP.
Kyle, Kristopher & Scott Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing 34*. 513–535.
Linacre, Michael. 2009. *A user's guide to FACETS v 3.66.0*. Chicago: Winsteps.
Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics 15*. 474–496.
Lu, Xiaofei. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Quarterly 45. 36–62.
Martínez, Ana Lahuerta. 2018. Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing 35*. 1–11.
McCarthy, Philip, Rebekah Guess & Danielle McNamara. 2009. The components of paraphrase evaluations. *Behavioural Research Methods 41*. 682–690.
McNamara, Danielle, Arthur Graesser, Philip McCarthy & Zhiqiang Cai, Z. 2014. *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge: CUP.
McNamara, Tim & Ute Knoch. 2012. The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing 29*. 555–576.
McNamara, Danielle, Arthur Graesser, Philip McCarthy & Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: CUP.
Ortega, Lourdes. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics 24.* 492–518.
van den Berg, Huub, Sven De Maeyer, Daphne van Weijen & Marion Tillema. 2012. Generalizability of text quality scores. In Elke Van Steendam, Marion Tillema, Gert Rijlaarsdam & Huub van den Bergh (eds.), 23–32. Leiden: Brill.
Verspoor, Marjolijn, Monika Schmid & Xiaoyan Xu. 2012. A dynamic usage-based perspective on L2 writing. *Journal of Second Language Writing 21*. 239–263.
Wiśniewski, Katrin**.** 2017. Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning 67*. 232–253.
Wolfe-Quintero, Kate, Shunji Inagaki & Hae-Young Kim. 1998. *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawaii Press.

Appendix 1

Table A. Descriptive statistics for the count variables across CEFR levels: grade 8

Index (Number of …)	A1 (n=37)		A2 (n=87)		B1 (n=70)		B2 (n=8)	
	M	SD	M	SD	M	SD	M	SD
Words	29.11	14.14	55.98	14.77	72.46	13.25	88.81	16.98
Sentences	3.44	1.60	5.29	1.85	6.18	1.76	7.56	1.52
Clauses	5.08	2.47	9.34	2.69	11.42	2.44	13.31	2.83
T-Units	3.54	1.88	5.67	1.93	6.86	1.72	8.40	1.28
Verb Phrases	5.33	2.83	10.04	2.83	12.75	2.72	15.35	2.34
Dependent clauses	1.31	0.77	2.89	1.17	3.68	1.31	4.17	1.57
Complex T-units	1.17	0.69	2.24	0.93	2.80	0.91	3.46	1.14
Coordinate phrases	0.45	0.43	0.97	0.68	1.00	0.68	0.90	0.67
Complex nominals	1.72	1.18	3.20	1.21	4.50	1.67	5.71	3.34

Table B. Descriptive statistics for the count variables across CEFR levels: Gymnasium

Index (Number of …)	A2 (n=31)		B1 (n=125)		B2 (n=39)	
	M	SD	M	SD	M	SD
Words	70.19	20.16	87.16	16.65	110.57	22.73
Sentences	5.85	1.94	6.62	1.67	8.18	4.67
Clauses	10.15	3.11	11.67	2.66	13.34	2.93
T-Units	6.62	2.29	7.65	1.96	8.56	1.75
Verb Phrases	11.6	3.81	14.25	3.22	17.40	3.72
Dependent clauses	3.59	1.62	4.18	1.65	5.01	1.87
Complex T-units	2.78	1.16	3.22	1.08	3.96	1.23
Coordinate Phrases	1.45	0.99	1.79	0.87	2.27	1.24
Complex nominals	6.44	2.33	7.59	1.91	10.52	3.12

Table C. Descriptive statistics for the syntactic complexity indices from L2SCA across CEFR levels: grade 8

Index	**A1 (n=37)**		**A2 (n=87)**		**B1 (n=70)**		**B2 (n=8)**	
	M	SD	M	SD	M	SD	M	SD
Sentence length	8.58	2.01	11.49	2.95	12.66	3.08	12.71	4.49
T-unit length	8.53	2.35	11.08	2.97	11.42	2.43	11.02	3.11
Clause length	5.97	1.44	6.38	0.92	6.59	0.81	6.99	0.79
Clauses per sentence	1.52	0.48	1.88	0.51	1.98	0.57	1.89	0.62
Clauses per T-unit	1.44	0.38	1.77	0.47	1.76	0.41	1.62	0.41
Complex T-units per T-unit	0.35	0.21	0.44	0.20	0.44	0.17	0.43	0.19
Dependent clauses per clause	0.24	0.14	0.30	0.11	0.32	0.09	0.30	0.08
Dependent clauses per T-unit	0.45	0.34	0.62	0.36	0.63	0.29	0.52	0.22
Coordinate phrases per clause	0.09	0.10	0.12	0.10	0.09	0.07	0.07	0.06
Coordinate phrases per T-unit	0.13	0.14	0.22	0.18	0.18	0.15	0.11	0.09
T-units per sentence	1.01	0.22	1.06	0.15	1.12	0.15	1.13	0.13
Complex nominals per clause	0.41	0.22	0.41	0.16	0.43	0.16	0.42	0.22
Complex nominals per T-unit	0.60	0.38	0.74	0.37	0.80	0.37	0.73	0.48
Verb phrases per T-unit	1.52	0.45	1.92	0.54	1.97	0.43	1.88	0.44

Table D. Descriptive statistics for the syntactic complexity indices from L2SCA across CEFR levels: Gymnasium

Index	**A2 (n=31)**		**B1 (n=125)**		**B2 (n=39)**	
	M	SD	M	SD	M	SD
Sentence length	13.06	3.59	14.24	3.78	15.44	2.49
T-unit length	12.01	3.14	12.73	3.45	13.88	2.17
Clause length	7.42	1.16	7.91	1.32	8.91	1.75
Clauses per sentence	1.84	0.47	1.88	0.48	1.89	0.49
Clauses per T-unit	1.67	0.39	1.66	0.41	1.65	0.32
Complex T-units per T-unit	0.50	0.19	0.49	0.17	0.51	0.15
Dependent clauses per clause	0.34	0.12	0.37	0.11	0.40	0.14
Dependent clauses per T-unit	0.66	0.33	0.68	0.42	0.68	0.25
Coordinate phrases per clause	0.16	0.11	0.17	0.10	0.18	0.10
Coordinate phrases per T-unit	0.25	0.17	0.26	0.17	0.28	0.16
T-units per sentence	1.12	0.18	1.15	0.14	1.14	0.10
Complex nominals per clause	0.71	0.27	0.71	0.23	0.87	0.28
Complex nominals per T-unit	1.16	0.46	1.16	0.41	1.34	0.34
Verb phrases per T-unit	2.01	0.53	2.08	0.60	2.18	0.35

Table E. Descriptive statistics for the syntactic complexity indices from Coh-Metrix across CEFR levels: grade 8

Index	A1 (n=37)		A2 (n=87)		B1 (n=70)		B2 (n=8)	
	M	SD	M	SD	M	SD	M	SD
Sentence length (st.dev.)	2.80	1.81	5.06	2.21	6.37	2.33	5.96	3.01
Syntactic simplicity (z-score)	0.57	0.92	0.29	0.76	0.21	0.73	0.46	1.21
Syntactic simplicity (percentile)	61.32	21.56	57.62	21.44	56.55	20.35	63.05	31.61
Left embeddedness	1.48	0.68	1.97	1.04	2.00	0.71	1.97	0.47
Modifiers per noun phrase	0.45	0.15	0.42	0.11	0.43	0.09	0.43	0.12
Minimal edit distance for parts of speech	0.36	0.26	0.53	0.20	0.63	0.11	0.64	0.14
Sentence syntax similarity (adjacent sentences)	0.14	0.08	0.11	0.05	0.10	0.04	0.11	0.04
Noun phrase density	379.86	59.27	371.30	34.22	353.00	25.53	356.43	26.83
Verb phrase density	229.27	45.49	251.37	30.78	258.22	25.07	255.82	20.95
Adverbial phrase density	23.43	31.90	28.74	13.82	33.52	15.16	38.54	18.25
Preposition phrase density	74.23	30.06	86.51	20.86	86.37	17.75	93.37	16.06
Negation density	37.80	18.59	32.42	14.14	32.05	12.41	27.96	9.63
Gerund density	8.05	13.68	6.57	8.48	11.08	8.75	7.82	6.66
Infinitive density	4.78	7.49	14.55	9.48	22.32	10.69	18.19	11.67

Table F. Descriptive Statistics for the Syntactic Complexity Indices from Coh-Metrix across CEFR levels: Gymnasium

Index	A2 (n=31)		B1 (n=125)		B2 (n=39)	
	M	SD	M	SD	M	SD
Sentence length (st.dev.)	5.18	1.74	6.75	3.61	7.64	2.34
Syntactic simplicity (z-score)	0.16	0.71	-0.13	0.69	-0.16	0.53
Syntactic simplicity (percentile)	55.18	19.98	46.05	20.31	44.42	17.24
Left embeddedness	2.35	1.11	2.44	0.84	3.09	1.02
Modifiers per noun phrase	0.51	0.11	0.56	0.12	0.66	0.15
Minimal edit distance for parts of speech	0.62	0.11	0.63	0.10	0.67	0.06
Sentence syntax similarity (adjacent sentences)	0.13	0.04	0.10	0.03	0.09	0.02
Noun phrase density	359.89	30.80	346.15	28.32	340.90	24.50
Verb phrase density	241.01	29.03	239.40	27.54	238.67	28.95
Adverbial phrase density	35.37	15.47	42.96	14.45	42.39	9.63
Preposition phrase density	82.78	27.61	91.37	18.29	97.13	17.86
Negation density	24.55	12.84	21.07	10.34	19.19	9.54
Gerund density	13.60	11.36	16.11	9.43	19.92	10.67
Infinitive density	15.96	13.93	17.20	9.67	20.52	10.65

III

SYNTACTIC COMPLEXITY IN YOUNG AND ADULT EFL LEARNERS' WRITING - DOES THE AGE MATTER WHEN THE LEARNERS' CEFR LEVEL IS THE SAME?

by

Khushik, G. A.

Submitted.

Request a copy from the author.