Ghulam Abbas Khushik* and Ari Huhta

Syntactic complexity in Finnish-background EFL learners' writing at CEFR levels A1–B2

https://doi.org/10.1515/eujal-2021-0011

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. Ar-

***Corresponding author: Mr. Ghulam Abbas Khushik,** University of Jyväskylä, center for applied language studies, Agora, Building Ag, 5th floor, Mattilanniemi 2, Jyväskylä,
E-Mail: ghabkhus@gmail.com
Prof. Ari Huhta, University of Jyväskylä, center for applied language studies, Jyväskylä,
E-Mail: ari.huhta@jyu.fi

viointiaineisto 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 levels 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; Huhta 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., Huhta et al., forthcoming). Furthermore, findings from Multidimensional studies on register variation in speaking and writ-

ing 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, Khushik and Huhta (2020) 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 (Khushik 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 (includ-

ing 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.

Table 2: (continued)

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.

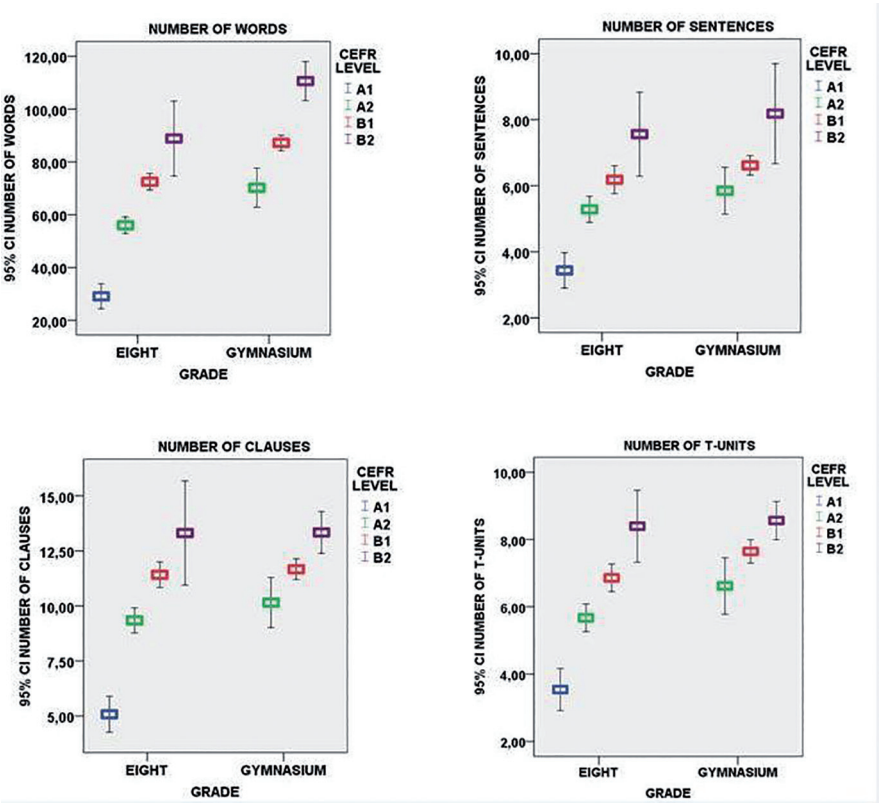

Figure 1: Error-bar charts for the essential count variables at different CEFR levels

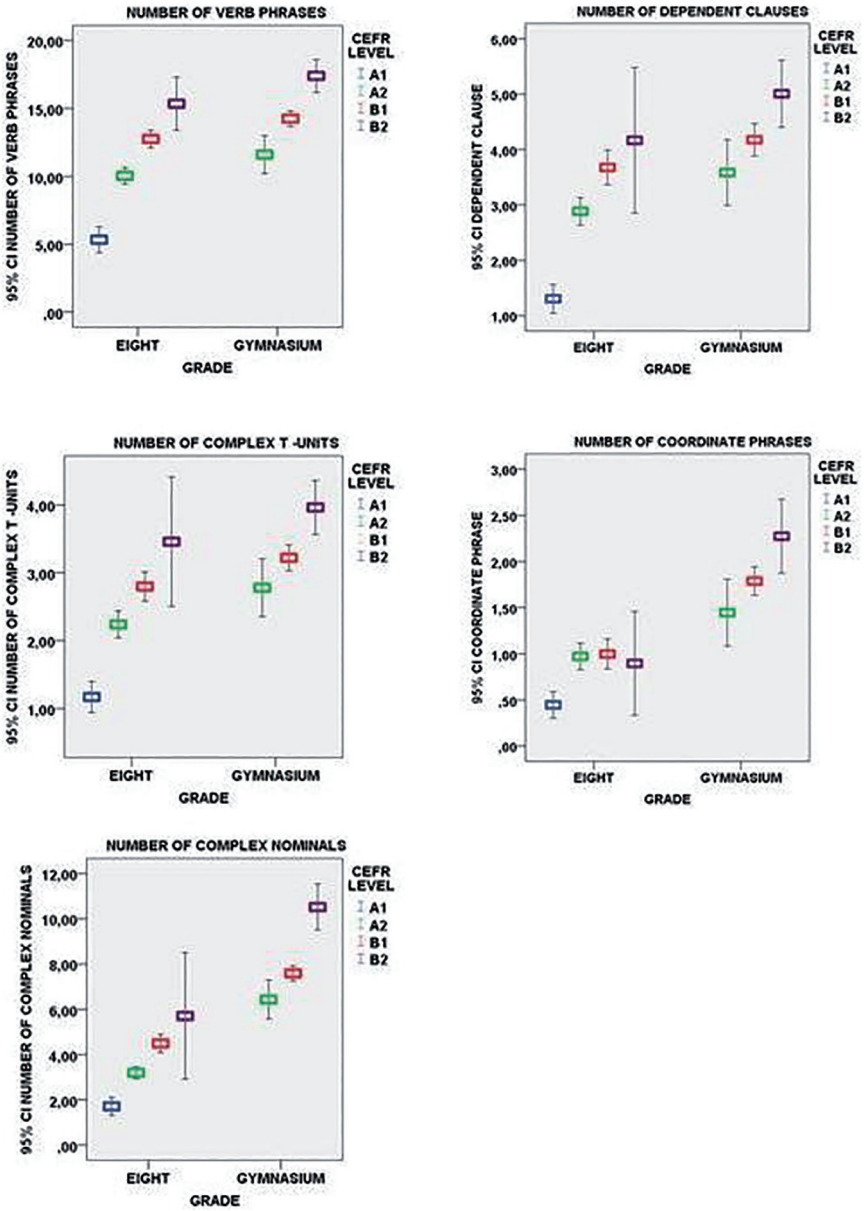

Figure 2: Error-bar charts for the more complex count variables at different CEFR levels

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

Table 6: (continued)

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 corre-

lated 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.

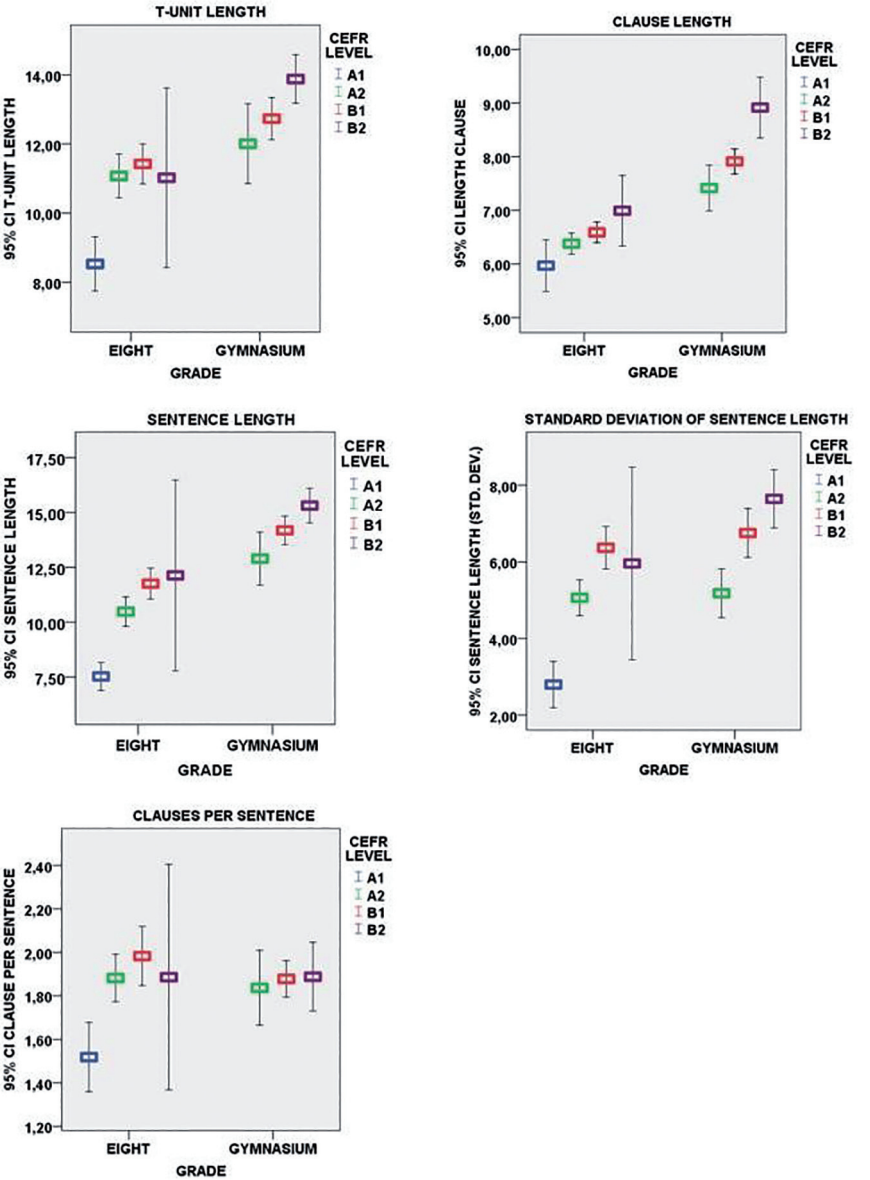

Figure 3: Error-bar charts for the mean clause, T-unit and sentence length, and clauses per sentence at different CEFR levels

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.

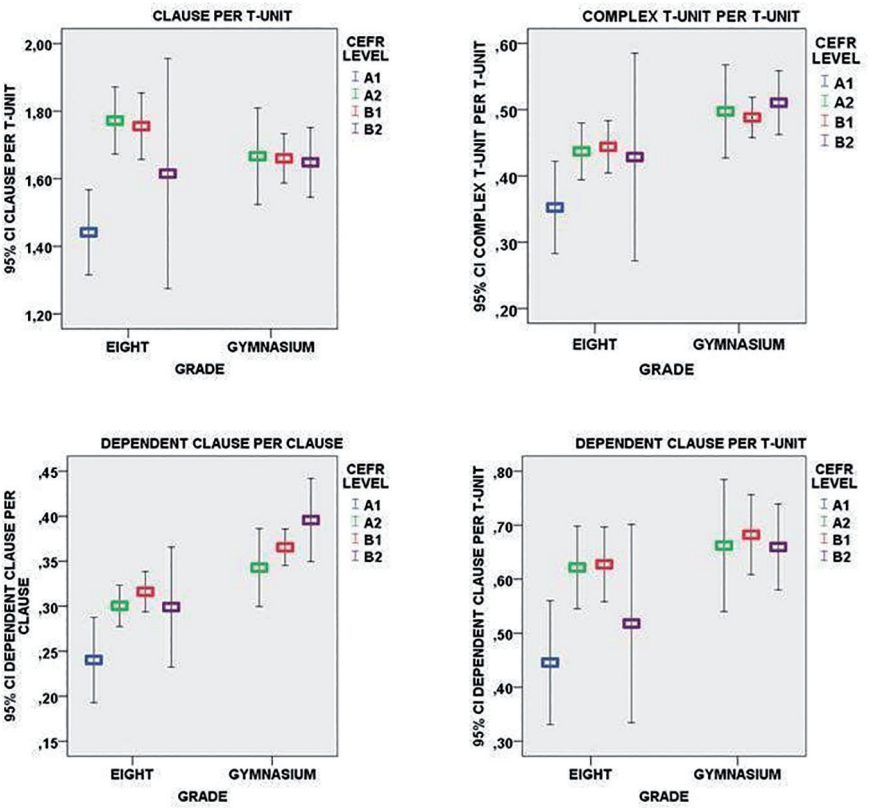

Figure 4: Error-bar charts for subordination indices at different CEFR levels

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

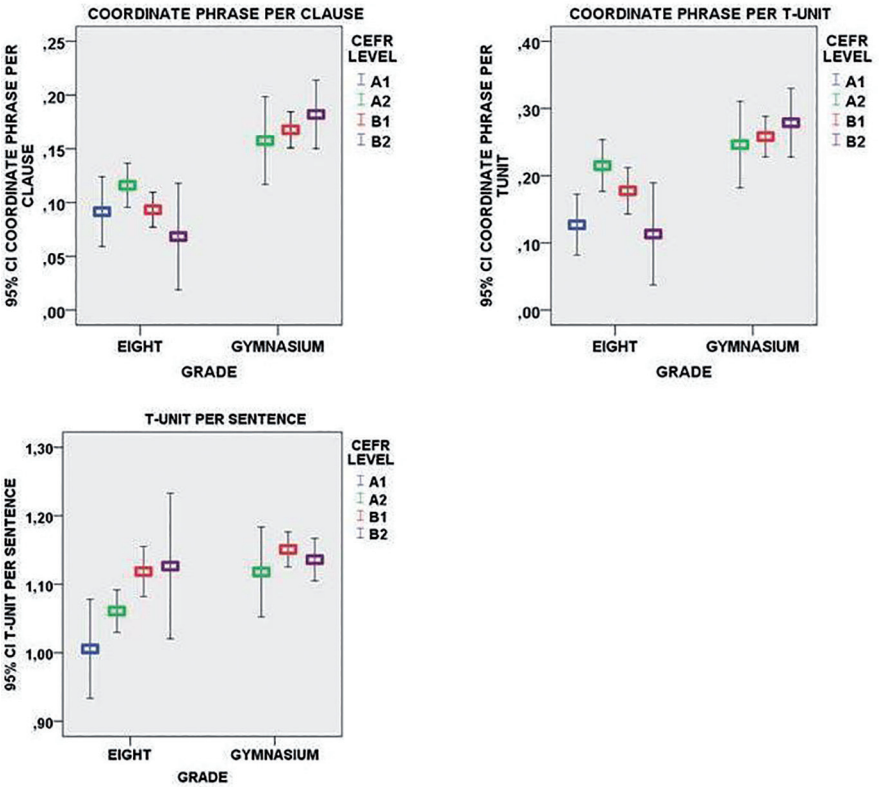

Figure 5: Error-bar charts for coordination indices at different CEFR levels

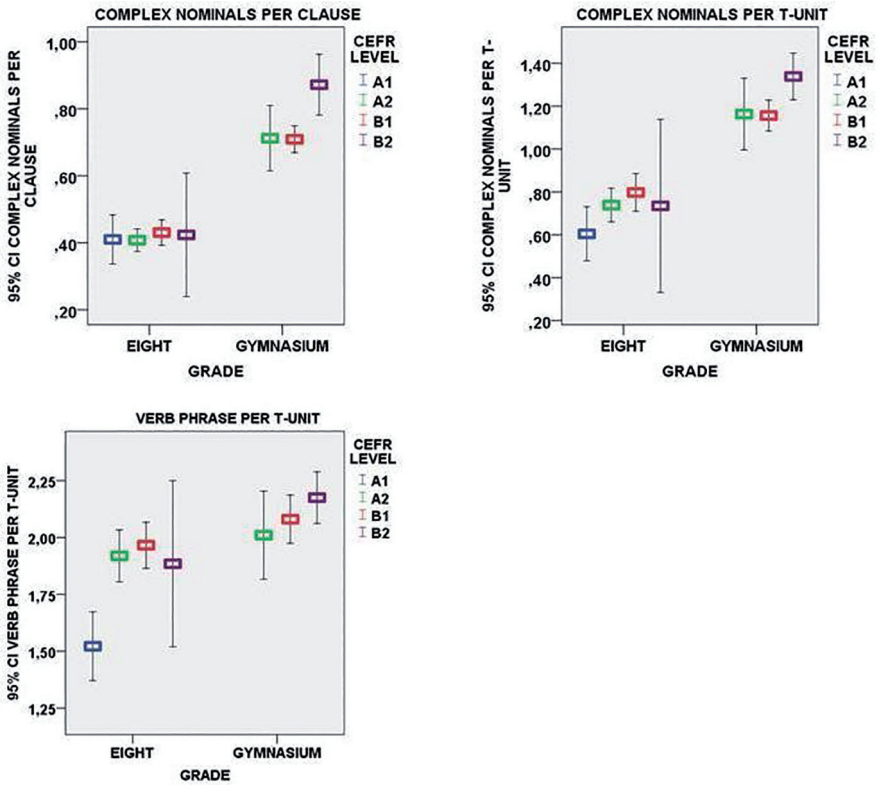

Figure 6: Error-bar charts for particular structures at different CEFR levels

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).

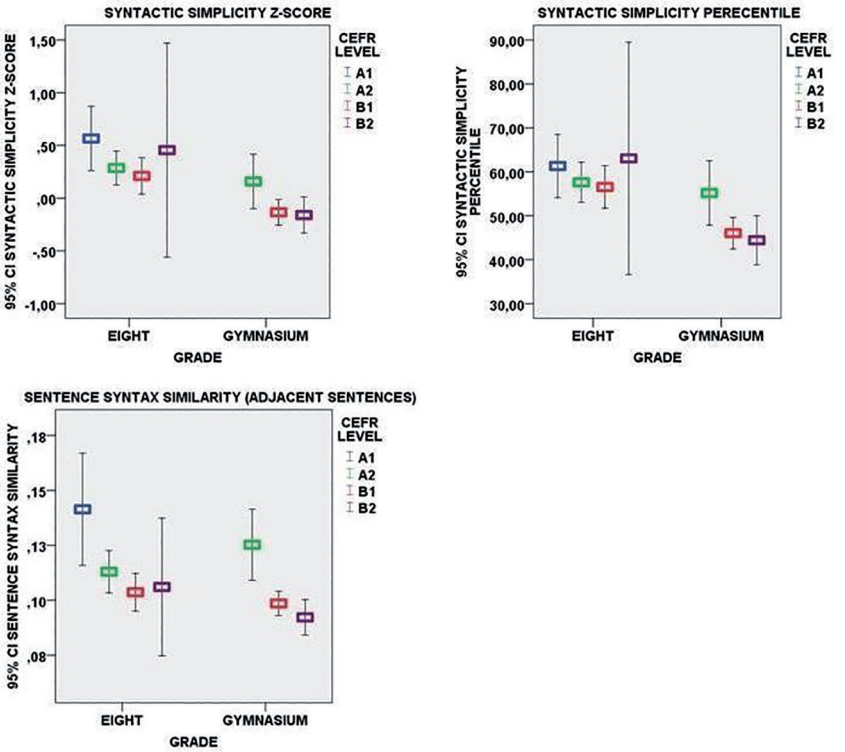

Figure 7: Error-bar charts for syntactic simplicity and sentence syntax similarity indices at different CEFR levels

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

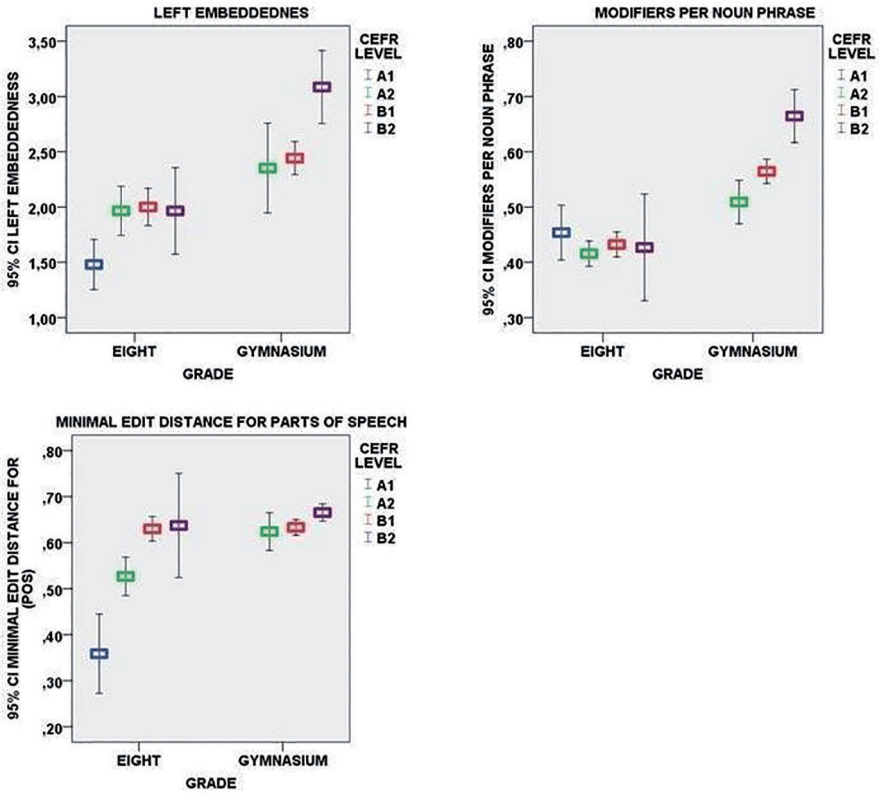

Figure 8: Error-bar charts for syntactic complexity indices from Coh-Metrix at different CEFR levels

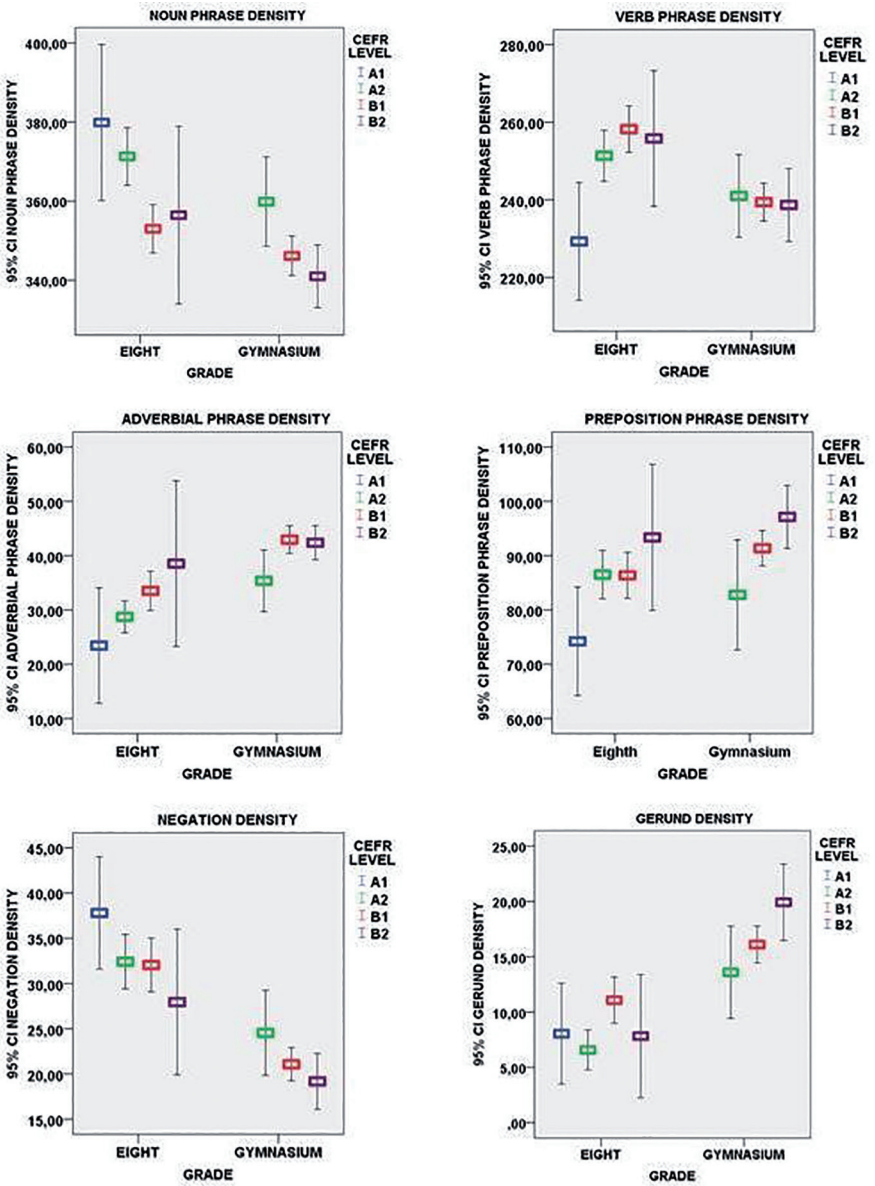

Figure 9: Error-bar charts for syntactic pattern density indices at different CEFR levels

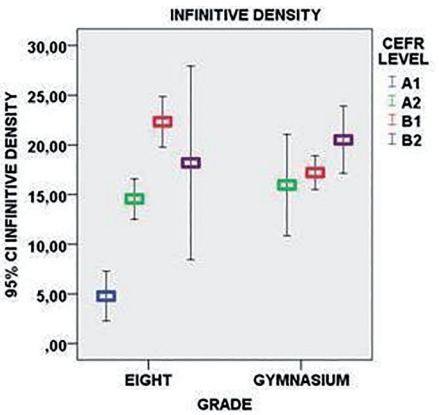

Figure 10: Error-bar chart for Infinitive Density (Syntactic pattern density) at different CEFR levels

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		

Table 14: (continued)

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, Khushik and Huhta (2020) 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 differ-

ent 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 Khushik and Huhta (2020) 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 (Huhta 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 ta-

ken 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 Khushik and Huhta (2020) 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.

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.

Khushik Ghulam Abbas, Ari Huhta, Investigating Syntactic Complexity in EFL Learners' Writing across Common European Framework of Reference Levels A1, A2, and B1, Applied Linguistics, Volume 41, Issue 4, August 2020, Pages 506–532, https://doi.org/10.1093/applin/amy064

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 wri ters 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. 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