Investigating Syntactic Complexity in EFL learners' writing across Common European Framework of Reference Levels A1, A2, and B1

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 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 3rd 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 NEAR HERE]

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 both 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 NEAR HERE]

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 NEAR HERE]

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 NEAR HERE]

[FIGURE 3 NEAR HERE]

[FIGURE 4 NEAR HERE]

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 NEAR HERE]

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 separataion. 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 NEAR HERE]

[FIGURE 6b NEAR HERE]

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 NEAR HERE]

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.

FIGURES

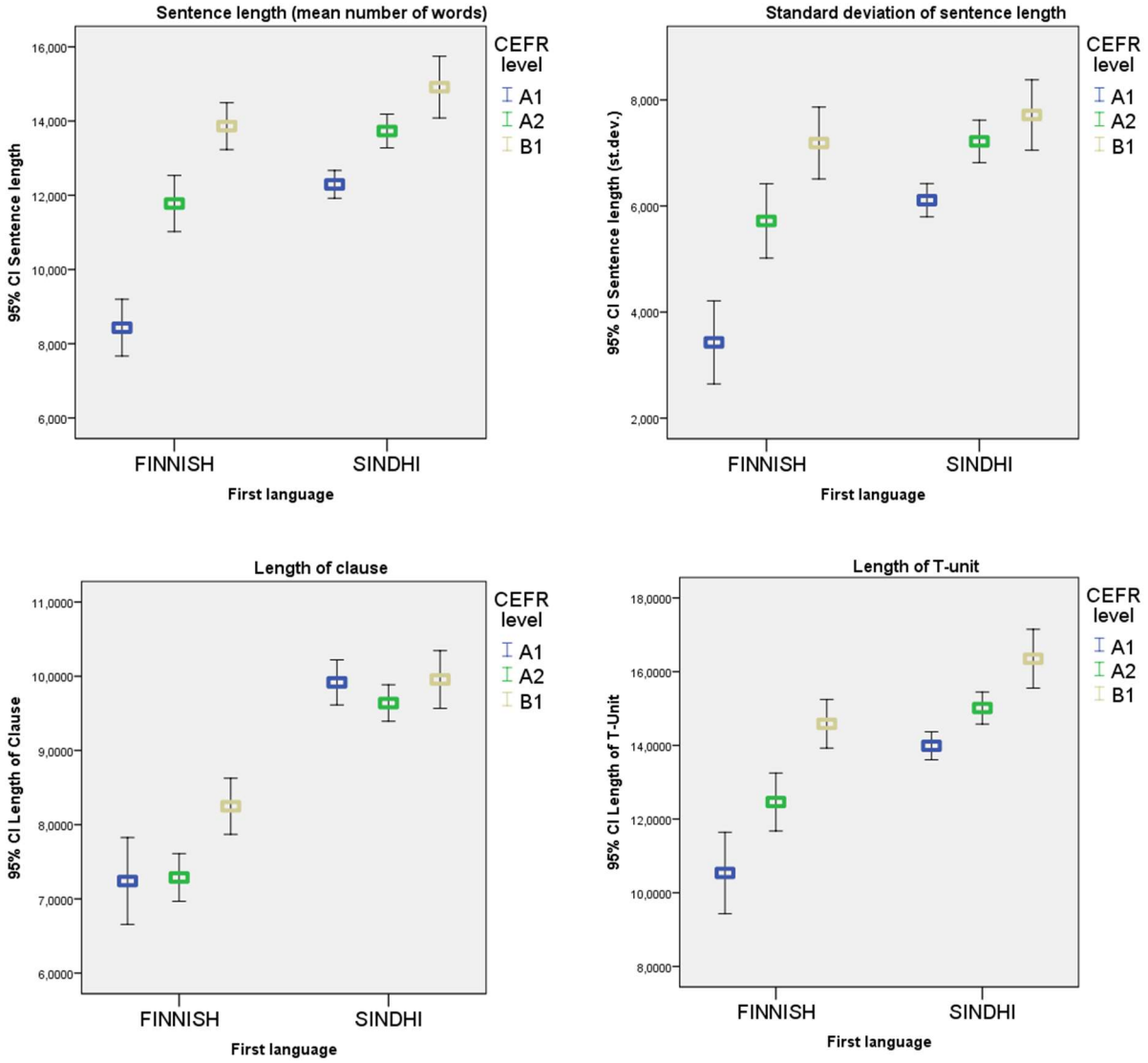


Figure 1. Error-bar charts for differences in the length of production units

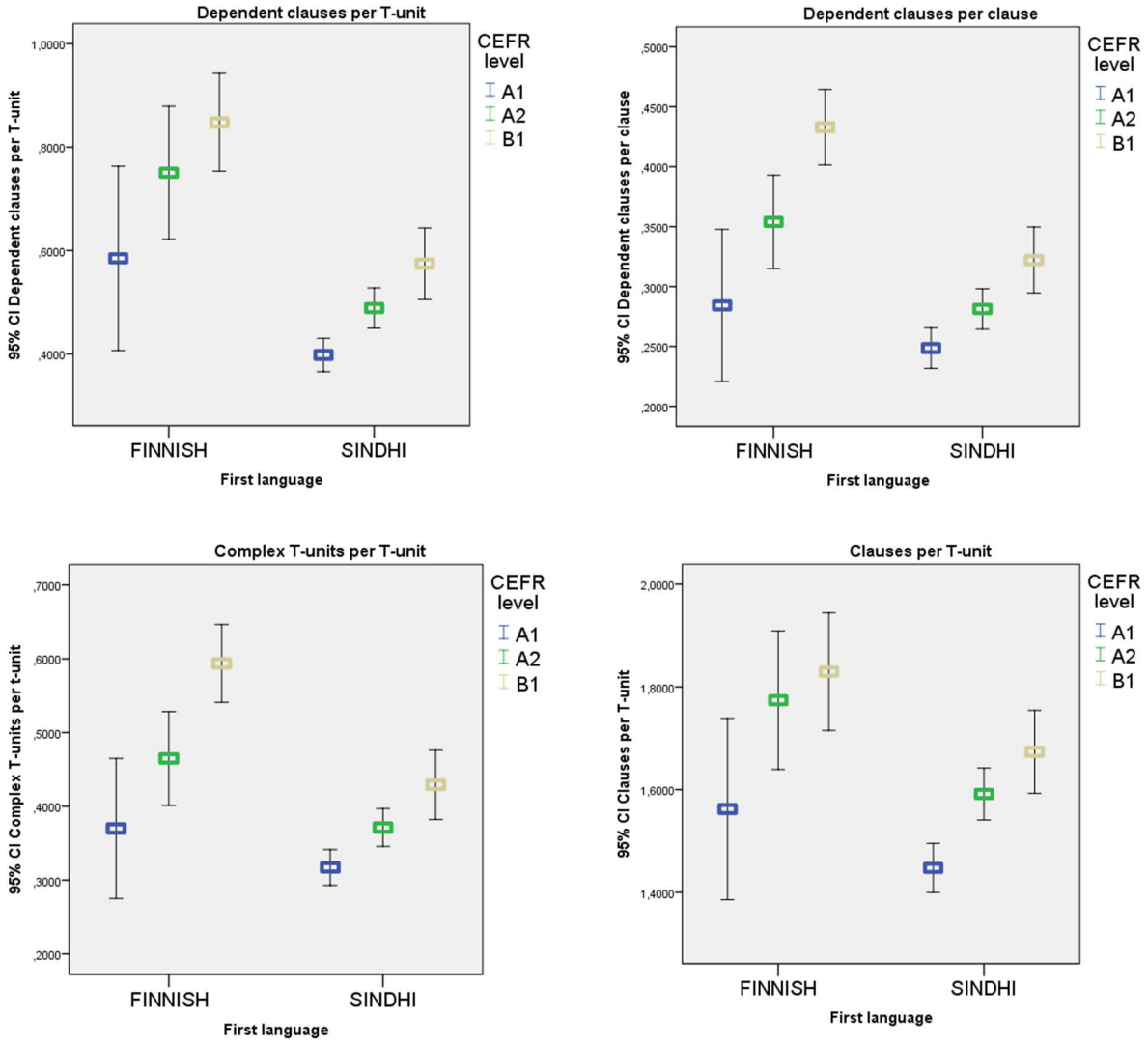


Figure 2. Error-bar charts for differences in subordination

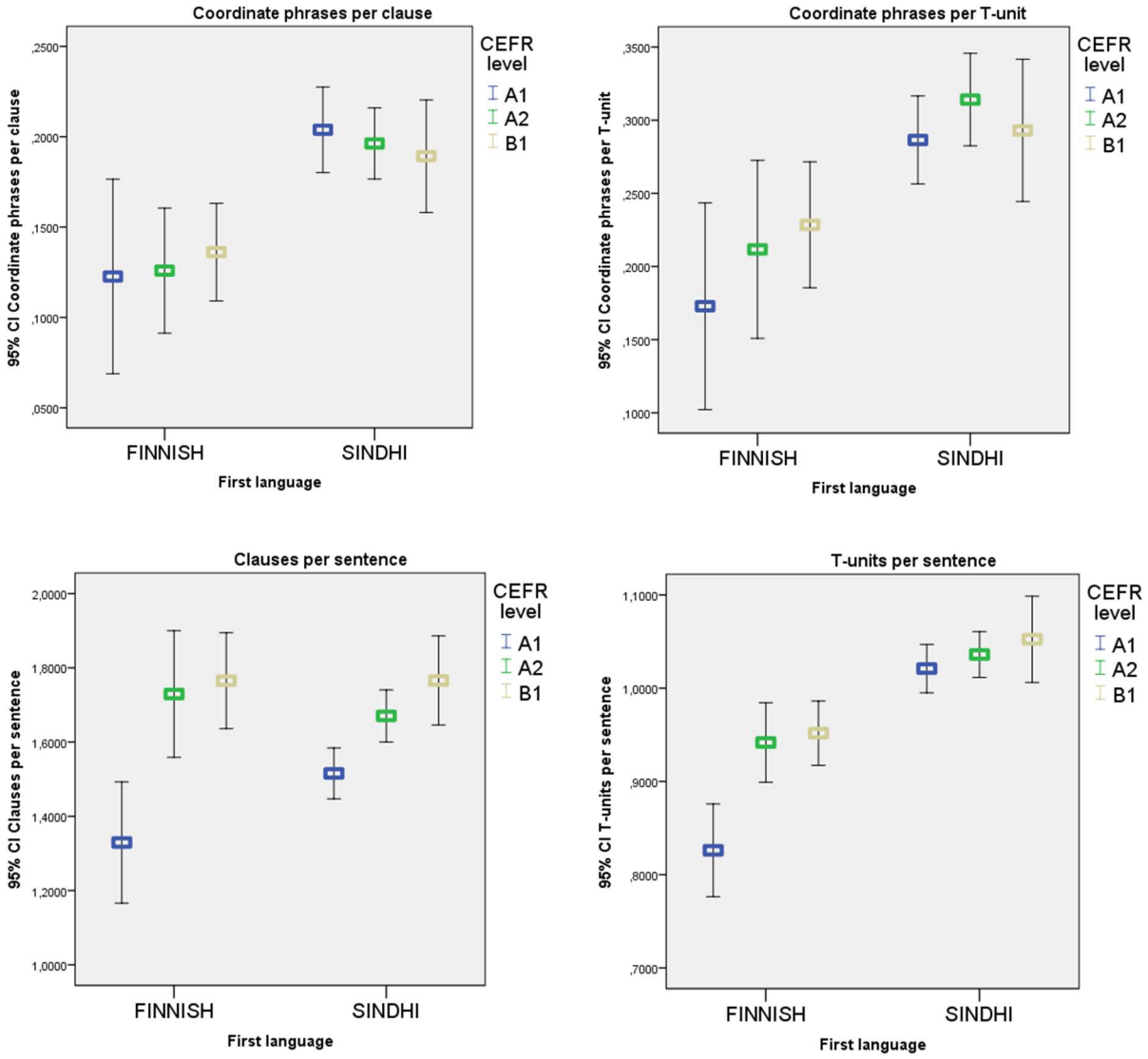


Figure 3. Error-bar charts for differences in coordination

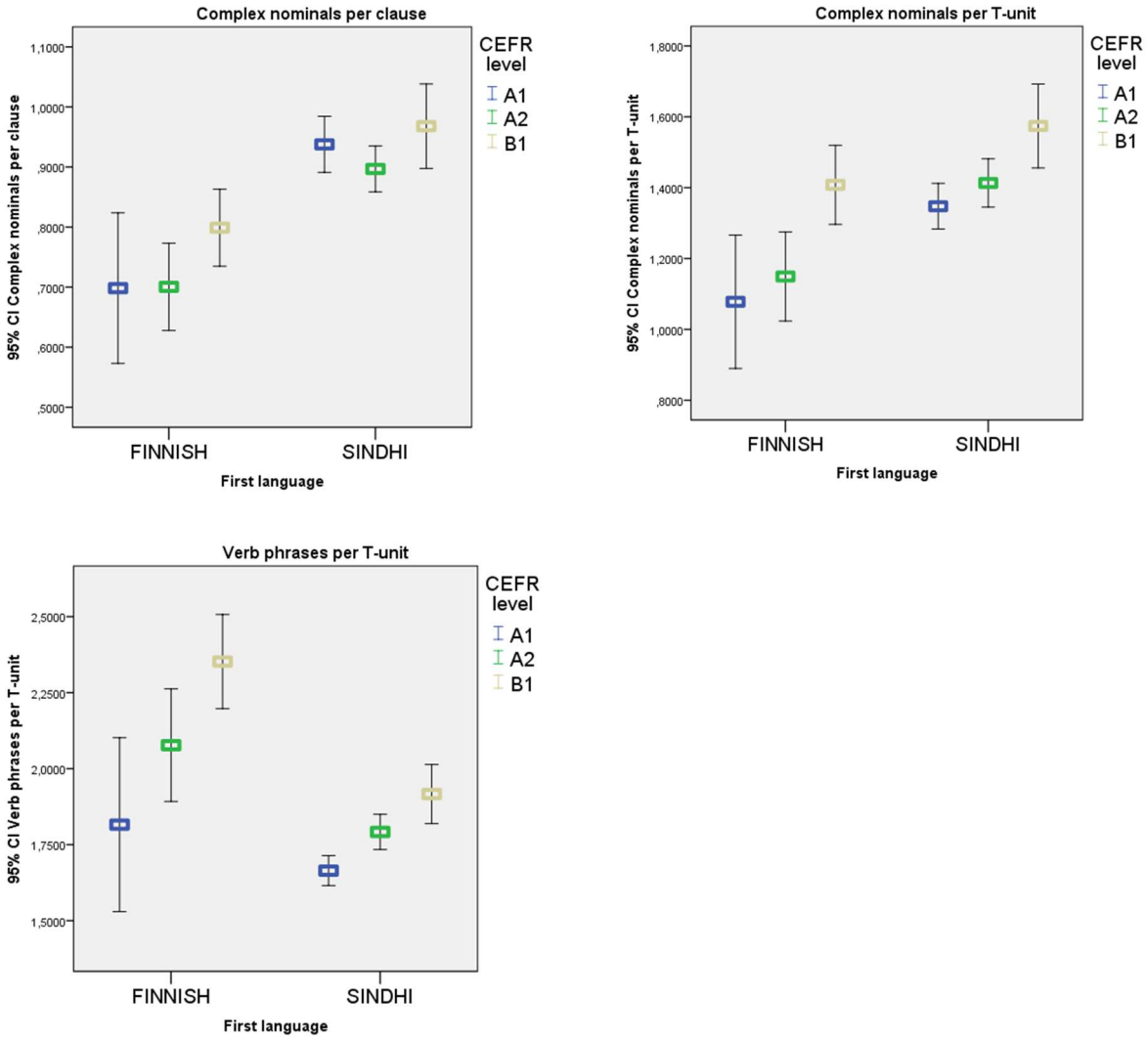
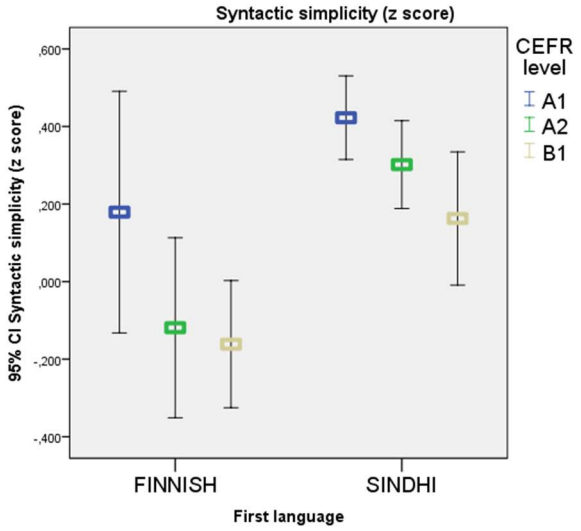
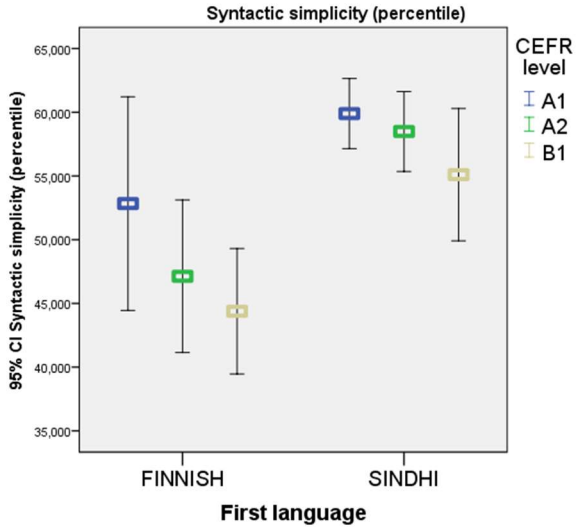
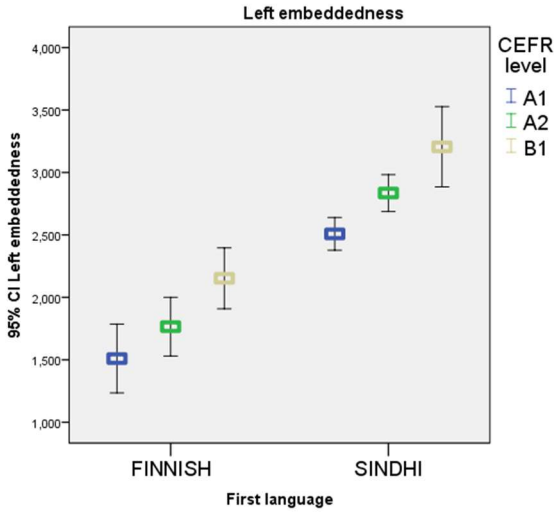
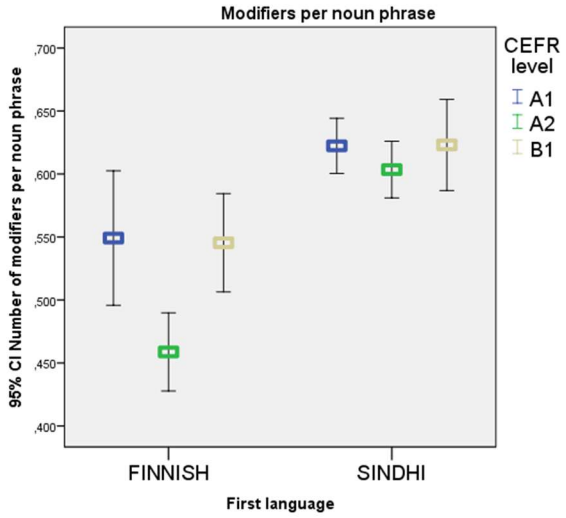


Figure 4. Error-bar charts for differences in phrasal sophistication


Syntactic simplicity (z score)
95% CI Syntactic simplicity (z score)
,600
,400
,200
,000
-,200
-,400
FINNISH
SINDHI
First language
CEFR level
A1
A2
B1


Syntactic simplicity (percentile)
95% CI Syntactic simplicity (percentile)
65,000
60,000
55,000
50,000
45,000
40,000
35,000
FINNISH
SINDHI
First language
CEFR level
A1
A2
B1


Left embeddedness
95% CI Left embeddedness
4,000
3,500
3,000
2,500
2,000
1,500
1,000
FINNISH
SINDHI
First language
CEFR level
A1
A2
B1


Modifiers per noun phrase
95% CI Number of modifiers per noun phrase
,700
,650
,600
,550
,500
,450
,400
FINNISH
SINDHI
First language
CEFR level
A1
A2
B1

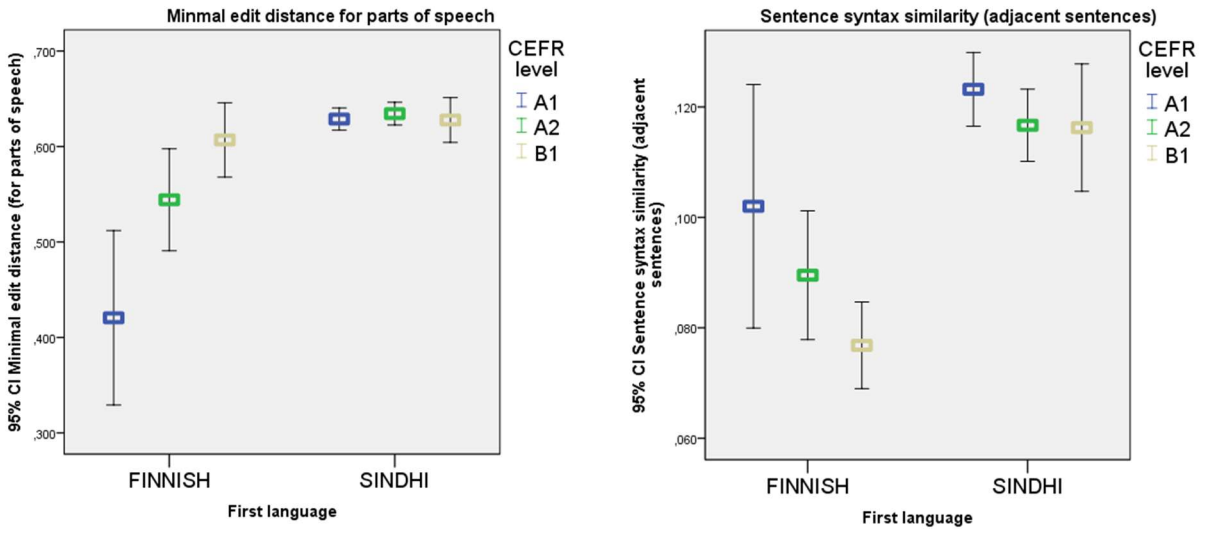


Figure 5. Error-bar charts for differences in working memory load, referencing expressions and syntactic variety and simplicity

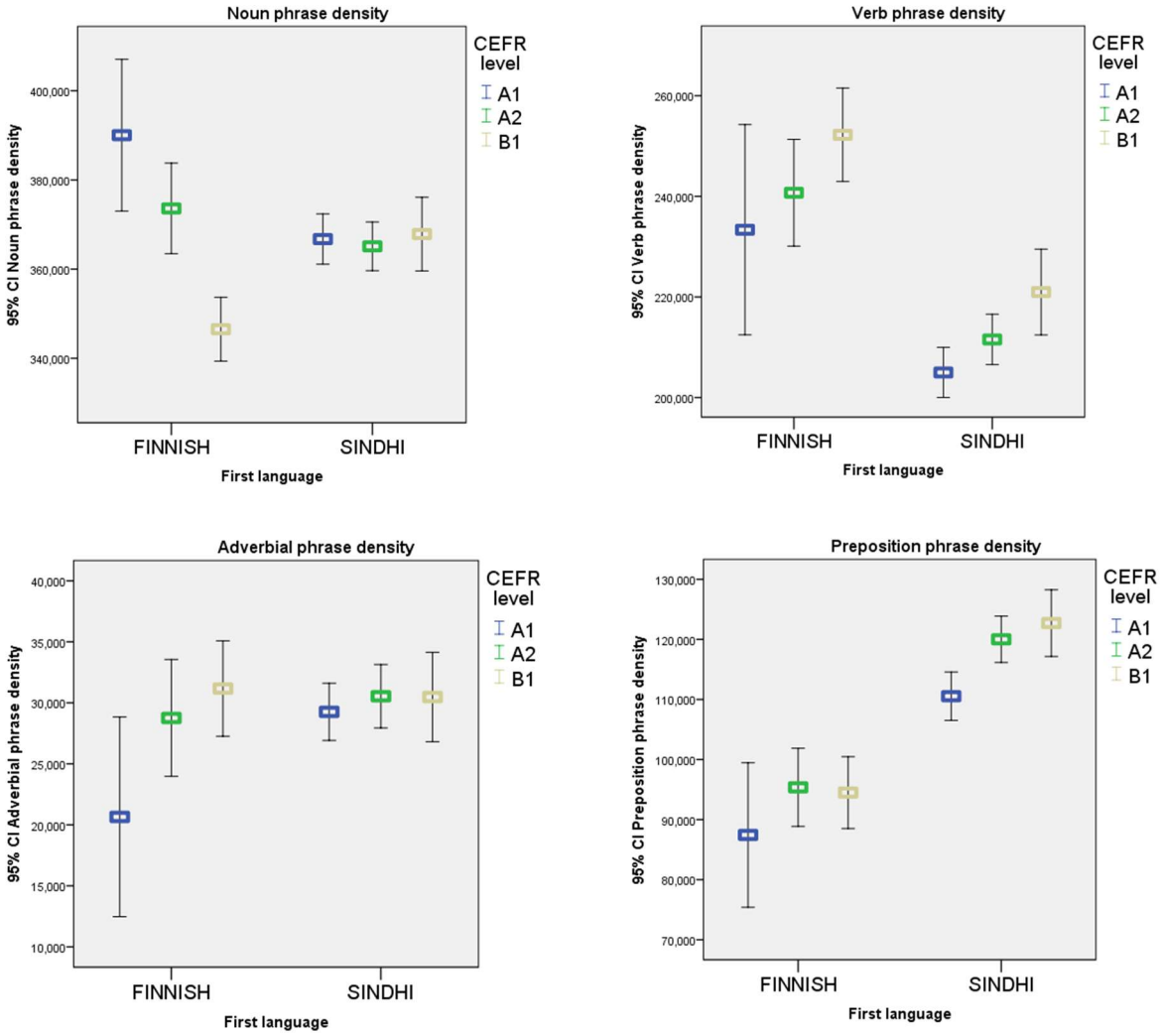


Figure 6a. Error-bar charts for differences in phrasal density

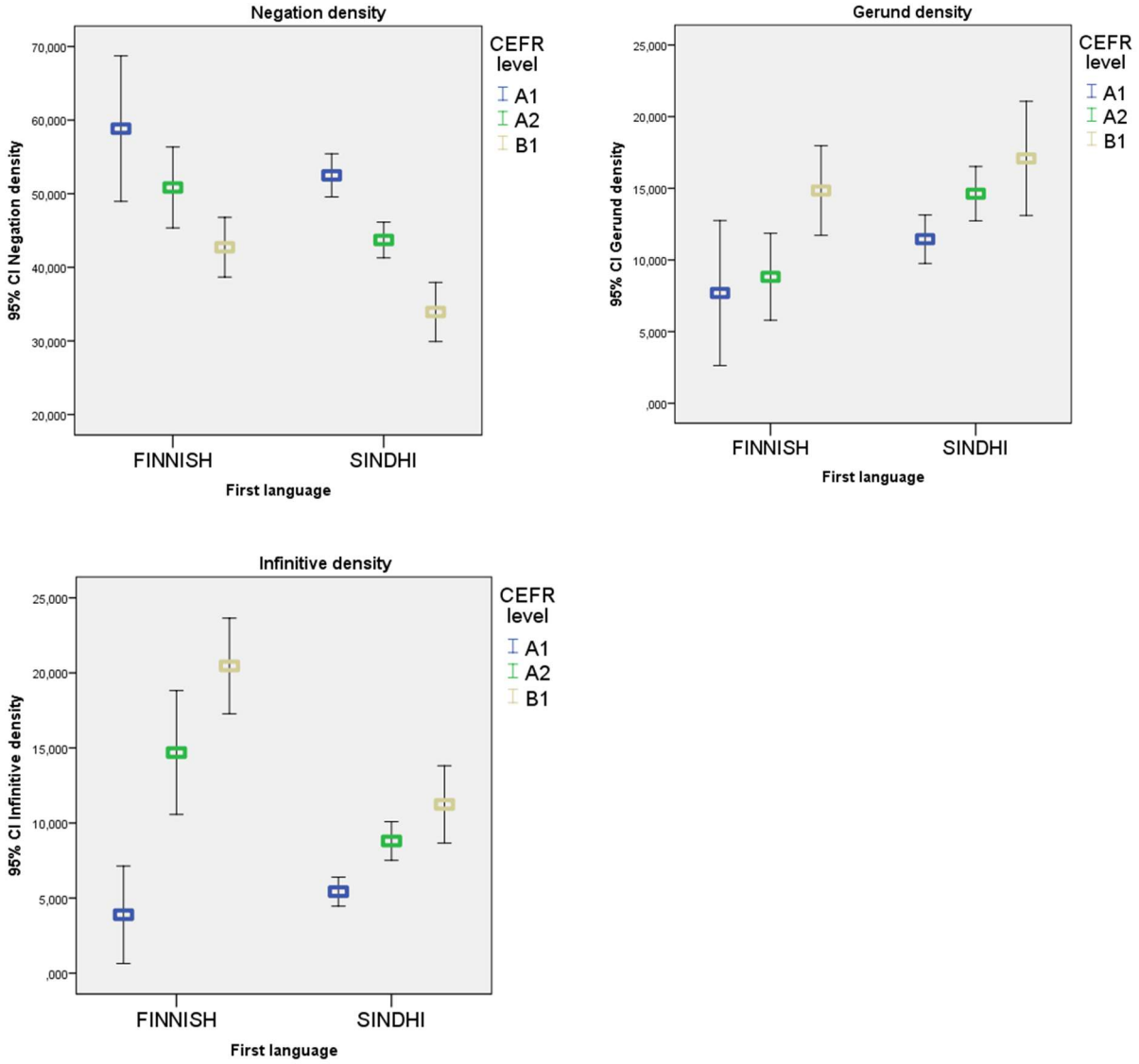


Figure 6b. Error-bar charts for differences in phrasal density

TABLES

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	A2 vs B1

clause ratios; noun phrases per clause

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%)

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
	Clause per sentence	Adverbial phrase density
Dependent clauses per T-unit	Adverbial phrase density	Gerund density
Complex T-unit per t-unit	(Verb phrases per T-unit)	

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)