Apples – Journal of Applied Language Studies
Vol. 14, 2, 2020, 67–92

Measuring syntactic complexity in learner Finnish

Taina Mylläri, University of Jyväskylä

In the study of complexity, accuracy and fluency (CAF), syntactic complexity can be measured by a multitude of measures. Traditionally, the measures are quantitative and they use production units such as words, clauses, T-units, and sentences. Despite the vast number of measures available, many studies have used only one or two of them, or parallel ones tapping the same component of complexity. The present study explores syntactic complexity using seven frequently used quantitative complexity measures to gauge different facets of complexity in written learner Finnish. The data of the study consist of texts written by adult and adolescent language learners, and they cover proficiency levels from beginner (A1) to advanced learner (C2) in the Common European Framework of Reference (CEFR). According to the results, changes in the measures are not linear from one proficiency level to the next. The results also show that while all the selected measures catch some statistically significant differences between proficiency levels in adult language learner texts, only four measures do so in adolescent language learner texts. The results also suggest that the measures are sensitive to task type.

Keywords: L2 writing, complexity, learner Finnish

1 Introduction

Learner language complexity can be defined as "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003, p. 492). Following Norris and Ortega (2009), complexity is most often considered a multi-faceted concept, and there is a multitude of quantitative measures of syntactic complexity available. Among the most popular are length-based measures, such as mean length of T-unit and mean length of clause, measures based on subordination, such as mean number of clauses per T-unit or mean number of dependent clauses per clause, and measures based on features considered sophisticated (Bulté & Housen, 2012; Norris & Ortega, 2009; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). Nevertheless, many studies use only one or two measures, or they use measures tapping the same aspect of complexity (Bulté & Housen, 2012, p. 34) or small datasets (Lu, 2011). This makes comparisons between studies difficult (e.g. Ellis & Barkhuizen, 2005; Pallotti, 2015).

According to a research synthesis by Ortega (2003), the most frequent measures of syntactic complexity have been mean length of clause (MLC), mean length of sentence (MLS), mean length of T-unit (MLTU), mean number of T-units per

Corresponding author's email: taina.myllari@jyu.fi
eISSN: 1457-9863
Publisher: University of Jyväskylä, Language Campus

https://apples.journal.fi
https://doi.org/10.47862/apples.99134

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

sentence (TU/S), mean number of clauses per T-unit (C/TU), and mean number of dependent clauses per clause (DC/C). Previous studies using these measures have yielded inconsistent results on the development of complexity across time or proficiency levels (e.g. Housen, De Clercq, Kuiken, & Vedder, 2019; Ortega, 2003; Wolfe-Quintero et al., 1998). Recent studies also indicate that there are differences in the development of complexity between languages (e.g. Gyllstad, Granfeldt, Bernardini, & Källkvist, 2014; Kuiken & Vedder, 2019).

In the present study, these measures are applied to written learner Finnish, and syntactic complexity is measured across the proficiency levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). The aim is on the one hand to test how traditional quantitative measures gauge syntactic complexity on different proficiency levels in a language that is structurally different from those more frequently studied, and on the other hand to examine if these measures could be used to indicate learner proficiency in Finnish. Seven quantitative measures, chosen on the basis of prior research on other languages, are used to tap different dimensions of syntactic complexity. The research questions are:

RQ1. How does syntactic complexity in written learner Finnish develop across CEFR proficiency levels when measured quantitatively?
RQ2. How well do the quantitative measures used in this study differentiate CEFR proficiency levels in written learner Finnish?

To answer the research questions, a pseudo-longitudinal learner Finnish corpus of texts written by second language (L2) learners with various backgrounds is analysed. Two different age groups of learners, i.e. adults and adolescents (12 to 16 years of age), are included to test if the measures yield the same results in both age groups. The adult L2 learners' texts cover all CEFR levels, from A1 to C2, and the adolescent L2 learners' texts cover levels A1 to B1/B2. The texts have been elicited with communicative tasks and divided into three types, according to task: informal messages, formal messages, and argumentative texts. These three types are analysed separately to see if there are differences in the results between the text types.

The theoretical background and measures used in the present study are introduced in Section 2. The data and methods are introduced in Section 3. The results for each measure are reported in Section 4. The article ends with discussion and conclusions in Section 5.

2 Measuring syntactic complexity

Previous studies have provided inconsistent results on the development of complexity (e.g. Housen et al., 2019). This may depend on the multitude of ways complexity has been operationalised and measured in the studies (Housen et al., 2019; Housen, Kuiken, & Vedder, 2012; Norris & Ortega, 2009; Pallotti, 2015), but differences in the settings of the studies (e.g. Ortega, 2003), individual variation in learner language (e.g. Larsen-Freeman, 2006), the non-linear development of language as a complex system (e.g. Larsen-Freeman, 2009), or complexity being manifested in different ways at different developmental stages (Housen et al., 2019; Norris & Ortega, 2009) may also partly explain the differences. Some of the inconsistencies may also be caused by typological differences between languages (Housen et al. 2019; Bernardini & Granfeldt, 2019).

One possible source of the mixed results are also the challenges involved in coding learner language. The most frequently used quantitative measures of syntactic complexity rely on production units such as clauses, sentences, and T-units, but these units can be ambiguous in both spoken and written learner language (e.g. Foster, Tonkyn, & Wiggelsworth, 2000; Martin, 2013). Differences in the definitions used in segmenting the data (e.g. Bulté & Housen, 2012; Wolfe-Quintero et al., 1998) or different interpretations made by annotators (e.g. Lu, 2010; Martin, 2013) may lead to differences in the number or length of the production units used. Such differences can affect the quantitative measures. For example, counting segments containing a non-finite verb form as either a dependent clause or as a part of a verb structure within a clause is likely to affect the number of words per clause (e.g. Bulté & Housen, 2012, pp. 39–40), with the number of dependent clauses and the total number of clauses then also affecting all the measures using these production units as units of measure (see also Mylläri, 2020).

Norris and Ortega (2009) suggest that syntactic complexity should be studied as a multi-faceted construct, and that different dimensions of complexity–overall complexity, complexity via subordination, subclausal complexity and, especially on lower proficiency levels, complexity via coordination–should be taken into account when measuring it. In addition to measures based on length or ratios of production units, complexity has been measured using the frequency of linguistic features that are considered sophisticated (e.g. Wolfe-Quintero et al., 1998).

Overall syntactic complexity, or general syntactic complexity, is typically measured by calculating mean length of sentence or T-unit in words (Bulté & Housen, 2012; Norris & Ortega, 2009). The results of studies using length-based measures have been mixed. In the research synthesis of Wolfe-Quintero et al. (1998), where sentence length (W/S) and T-unit length (W/T) were considered measures of fluency, they were both found to grow linearly with proficiency. According to Wolfe-Quintero et al. (1998), sentence length was found to correlate with proficiency in all the 10 studies, and T-unit length in 28 of the 40 studies included in the synthesis. Bulté and Housen (2012), however, point out that while length-based measures may show linear increase at some phase of development, they may well plateau at some level, in much the same way as L1 development of mean length of utterance has been shown to do.

Measures based on subordination have been among the measures most widely used in studies on syntactic complexity (e.g. Bulté & Housen, 2012). According to Wolfe-Quintero et al. (1998), clauses per T-unit (C/TU) and dependent clauses per clause (DC/C) are good indicators of proficiency as they seem to grow linearly with proficiency, although only some of the studies in their research synthesis found a correlation between the measures and proficiency. The relevance of subordination in measuring syntactic complexity especially in writing has later been questioned. Biber, Gray, and Poonpon (2011) argue that subordination is more typical of spoken language than of academic texts. Bulté and Housen (2012) note that subordination ratios only gauge one type of complexity and ignore others, such as clausal coordination or complexity at the phrasal level. Martin, Mustonen, Reiman, and Seilonen (2010) have also questioned using subordination as an indicator of learner Finnish complexity since, in Finnish, there are no apparent syntactic or morphological differences between subordinate clauses and main clauses, with the exception of relative clauses.

Measures based on subordination also overlook coordination as a part of complexity (e.g. Bardovi-Harlig, 1992). This could partly be explained by coordination being often associated with lower proficiency levels, as development is generally thought to proceed from coordination at beginning levels to subordination at intermediate levels and phrasal-level elaboration at advanced levels (e.g. Norris & Ortega, 2009). While Wolfe-Quintero et al. (1998) conclude that the coordination ratio of T-units per sentence (TU/S) has not been shown to be useful in L2 research, Norris and Ortega (2009) suggest that measures of coordination should also be included, especially in studies using data on lower proficiency levels.

A measure taking both coordination and subordination into account is number of clauses per sentence (C/S). Wolfe-Quintero et al. (1998) found that this measure had been used in only one study, and in that case the growth in the measure had been statistically significant for only a part of the study. Nevertheless, Lu (2010, 2011) has included C/S in the 14 measures in his L2 Syntactic Complexity Analyzer and labelled it a measure of sentence complexity.

Although mean length of clause (MLC) in terms of number of words can be considered a length-based measure, it is more often regarded as a measure of clausal or phrasal complexity than overall complexity, for the reason that it shows lengthening within a clause, thus indicating elaboration on the phrasal level (e.g. Norris & Ortega, 2009). Previous studies have shown MLC to correlate with proficiency: learners on the higher proficiency levels tend to use longer clauses than those on lower proficiency levels (e.g. Lu, 2011; Ortega, 2003). Wolfe-Quintero et al. (1998) considered clause length (W/C) to be a measure of fluency, and they found that it grew linearly with proficiency in all nine studies included in the synthesis, although the correlation was not statistically significant in all the studies. A similar measure, i.e. mean number of finite verbs per total number of words, has also been found to develop linearly over time (Verspoor, Lowie, Chan, & Vahtrick, 2017). Mean number of finite verbs per total number of words is the same as mean length of clause provided that each clause in the data contains a finite verb and all the words in the word count belong to a clause.

Syntactic complexity in L2 Finnish has so far been studied using small datasets or by analysing the development of some specific structures. Alisaari (2016), who used mean length of T-unit (MLTU) as a measure of fluency in written learner Finnish on CEFR level A2, found no significant development in MLTU in narrative texts written by 32 learners at the beginning and at the end of a four-week course. Tilma (2014) studied the development of complexity and accuracy in foreign and second language Finnish using written data collected from eight students over a period of nine months. Among the measures used in her study were average length of sentence and average length of clause in morphemes, both of which she found increased over time, although she found no statistically significant correlation between the indices and development on the group level. She concluded that the best measure of syntactic complexity in learner Finnish was average length of clause in morphemes. Spoelman and Verspoor (2010) studied learner Finnish complexity and accuracy in a DST framework. Using a longitudinal data set of 54 writing samples collected from one L2 Finnish learner over a period of three years, they found a non-linear increase in complexity, including the sentence complexity ratio, which was based on the average number of dependent clauses per text.

The development of linguistic features in relation to an increase in proficiency from one CEFR level to the next in written learner Finnish has been studied using the CEFLING project data, from which the data of the current study are also drawn. Seilonen (2013) studied the use of indirect references, Kajander (2013) the use of existential sentences, and Reiman (2011a, 2011b, 2014) transitive constructions. All of these linguistic features showed growth in frequency, variation and accuracy across proficiency levels. There were, however, differences in the use of these linguistic features between the adult and adolescent language learners and between task types. The results indicate that there is a leap in frequency between levels A2 and B1 in the adult learner data, whereas in the adolescent learner data a similar difference can already be found between levels A1 and A2. (Kajander, 2013; Reiman, 2014; Seilonen, 2013.)

To test the usability of the frequently used quantitative measures in written learner Finnish, the following seven measures were selected to tap the different dimensions of syntactic complexity (Table 1). To tap overall or general complexity, mean length of sentence (MLS) and mean length of T-unit (MLTU) were used. Following Lu (2010, 2011, 2017), mean number of clauses per sentence (C/S) was also calculated as a measure of overall sentence complexity. For complexity via subordination, two measures, mean number of clauses per T-unit (C/TU) and mean number of dependent clauses per clause (DC/C), were used. Complexity via coordination was measured with the mean number of T-units per sentence (TU/S). Sub-clausal complexity was measured with mean length of clause (MLC).

Table 1. Complexity measures used in the present study.

Label	Measure	Formula
MLS	Mean length of sentence	Total number of words / total number of sentences
MLTU	Mean length of T-unit	Total number of words / total number of T-units
MLC	Mean length of clause	Total number of words / total number of clauses
TU/S	Mean number of T-units per sentence	Total number of independent clauses / total number of sentences
C/S	Mean number of clauses per sentence	Total number of clauses / total number of sentences
C/TU	Mean number of clauses per T-unit	Total number of clauses / total number of independent clauses
DC/C	Mean number of dependent clauses per clause	Total number of dependent clauses / total number of clauses

Six of the measures, i.e. MLS, MLTU, MLC, TU/S, C/TU, and DC/C, are among the most frequently used in research on L2 complexity, according to Ortega (2003).

3 Data and methods

3.1 Data

The data used in the present study are drawn from the pseudo-longitudinal corpus of the Jyväskylä University CEFLING project and they comprise 667 texts (48,876 tokens) from adult L2 Finnish learners and 411 texts (16,590 tokens) from adolescent L2 learners. In the CEFLING project, the adult L2 learners' texts were selected from the National Certificates of Language Proficiency examination database, and the adolescent L2 learners' texts were collected from pupils in school years 7, 8 and 9 (between 12 and 16 years of age), together with similar texts from their native Finnish (L1) counterparts. The texts were elicited through communicative writing tasks, and they have been arranged into groups according to the type of task: informal messages (e.g. an email to a friend), formal messages (e.g. a complaint to an online store), and argumentative texts (e.g. a text expressing an opinion on a given topic, such as the use of mobile phones at school). The adolescent L2 learners and L1 writers also wrote a narrative text. The participants had a limited time in which to complete the writing tasks, and the use of aids, such as dictionaries, was not allowed. (Alanen, Huhta, & Tarnanen, 2010; Jantunen & Pirkola, 2015; Martin et al., 2010.) The L2 messages and argumentative texts are used in the present study.

In the CEFLING project, the L2 texts were assessed on the proficiency levels of the Common European Framework of Reference (CEFR), using scales based on the framework. The assessment was done by a team of trained raters, and each text was rated by three raters. (Alanen et al., 2010.) The reliability of the assessment has been shown by quantitative and qualitative analysis (Huhta, Alanen, Tarnanen, Martin, & Hirvelä, 2014). There are adult learner texts on all the CEFR proficiency levels, from A1 to C2. For the adolescent learners, the proficiency levels range from A1 to B1 in the argumentative texts and to B2 in the informal and formal messages. However, there are only a few adolescent L2 learner texts at level B2.

When annotating the data for the present study, echoes of task prompts and segments consisting of verbless greetings or contact information, such as email or street addresses and phone numbers, were excluded from the analysis (cf. Foster et al., 2000, pp. 370–371). Additionally, four whole texts were excluded during annotation: two adult L2 learners' texts containing only verbless greetings or list items, and two adolescent L2 learners' texts with inconsistencies in task type or writer identification information.

For the analysis, the L2 texts were organised according to the CEFR level. The two age groups of L2 learners, adult and adolescent learners, were studied separately, because earlier studies (Kajander, 2013; Reiman, 2014; Seilonen, 2013) using the same data have shown developmental differences between adult and adolescent learners. To control for task or genre effect (see e.g. Michel, 2017), the three task types were kept separate. The number of texts and words in the data are presented in Table 2.

Table 2. The number of texts and words and the average length of texts in words.

	Informal messages			Formal messages			Argumentative texts		
	Texts	**Words**	**Average length**	**Texts**	**Words**	**Average length**	**Texts**	**Words**	**Average length**
L2 adult learners									
A1	39	1,582	40.56	22	752	34.18	50	2,261	45.22
A2	39	1,533	39.31	27	1,494	55.33	37	2,272	61.41
B1	41	2,453	59.83	42	2,290	54.52	43	5,142	119.58
B2	39	2,206	56.56	34	1,983	58.32	35	4,166	119.03
C1	26	1,528	58.77	45	3,337	74.16	46	5,876	127.74
C2	14	970	69.29	58	5,152	88.83	30	3,879	129.30
Total	**198**	**10,272**	**51.88**	**228**	**15,008**	**65.82**	**241**	**23,596**	**97.91**
L2 adolescent learners									
A1	25	677	27.08	33	860	26.06	32	775	24.22
A2	79	2,879	36.44	40	1,489	37.23	39	1,589	40.74
B1	64	2,971	46.42	40	1,980	49.50	40	2,232	55.80
B2	12	677	56.42	7	461	65.86	-	-	-
Total	**180**	**7,204**	**40.02**	**120**	**4,790**	**39.92**	**111**	**4,596**	**41.41**

The data come from 868 different writers. Of the 481 adult learners, 40 wrote three texts each, 106 wrote two texts each, and 335 one text each. Of the 212 adolescent learners, 45 wrote three texts each, 109 two texts each, and 58 one text each. In the CEFLING project, each text was placed on a proficiency level independently (e.g. Martin et al., 2010).

3.2 Production units and segmenting the data

For the argumentative texts, a manually segmented corpus from an earlier study (Mylläri, 2020) was used. The informal and formal messages were first split into sentences, and each sentence was then split into clauses using the clause-splitting feature of the Finnish Dependency Parser (Haverinen et al., 2014)[1]. The segmentation was manually checked by the author to ensure that it was in line with the guidelines described below, and exceptions were considered case by case (see also Mylläri, 2020).

Words and sentences were segmented based on orthography. A segment was counted as a word if it contained at least one letter or number, or a symbol such as the euro sign (€), and if it was separated from other text by a space or other orthographic indicator, such as punctuation. This definition of word was considered to be feasible for this study, as in Finnish there are no articles and only a few prepositions, and compound words are generally written as one orthographic unit (e.g. *olohuone* 'a/the sitting room, *olohuoneessa* 'in a/the sitting room').

A sentence was defined as an orthographic unit ending with proper punctuation or, in the absence of punctuation, with an end-of-line character. Each sentence was annotated to contain at least one independent clause. This was applied also to sentences containing only one clause starting with a subordinator (see also Foster et al., pp. 2000: 336; Kalliokoski, 2006) and to sentences not containing a finite verb.

A clause was defined as a segment within a sentence clustered around a finite verb (cf. Lu, 2010; Wolfe-Quintero et al, 1998, p. 123). Clauses beginning with a subordinator or the surface-level ellipsis of a subordinator and having a main clause within the same sentence were annotated as dependent clauses. All other clauses were labelled as independent. Two clauses concatenated without any connectors were assumed to be coordinated with each other. Three types of exception were allowed in order to include all the words in the analysis and to maintain possibly intended subordination, even if there was no finite verb in the main clause or in the subordinate clause. First, as mentioned above, sentences not containing a finite verb were considered to contain at least one clause. Second, segments not containing a finite verb but functioning as a main clause to at least one subordinate clause within the same sentence, and not being subordinated to or coordinated with another clause, were considered independent clauses. Third, when there was a main clause within the same sentence, segments beginning with a subordinator were counted as subordinate clauses even if they did not contain a finite verb.

A T-unit was defined as consisting of one independent clause together with all the dependent clauses that are either directly or indirectly connected to it within the same sentence.

3.3 Statistical methods

All seven measures were calculated for each text. The results were rounded to two decimal places. Group mean and median, standard deviation, and interquartile range were calculated for each proficiency level for all the task types (informal message, formal message, argumentative text) separately for the adult and adolescent learners. All the texts in the data were included in the analysis, and outliers were considered to be occurrences of individual variation and therefore included in the calculations and statistical analyses. The descriptive statistics are presented in Tables A1–A7 in Appendix A.

Before making the statistical comparisons between proficiency levels, the data were visualised using boxplots and Q-Q plots. Since sample sizes varied, two tests were used to test the normality of distribution. The Shapiro-Wilk test was used for samples of 50 texts or fewer and the Kolmogorov-Smirnov test for samples larger than 50 texts. Both tests indicated violations of normality (68% of the samples of 50 texts or fewer and 24% of the samples over 50 texts). Homogeneity of variance was therefore tested with the Fligner-Killeen test (e.g. Gries, 2013, p. 229), which showed that the assumption was violated in around 31% of the comparisons.

Because of the differences in sample size and violations of assumptions of normality of distribution and homogeneity of variance, both parametric and non-parametric tests were used to test for differences between proficiency levels. For omnibus tests, one-way ANOVA and the Kruskal-Wallis test were used. For both tests, the cut-off point for statistical significance was set at $p < .05$. For effect size, adjusted R-squared in ANOVA was used with the following guidelines: > .01 small, > .06 medium, and > .14 large (see e.g. Larson-Hall 2010: 119). For complexity measures with statistically significant differences in ANOVA or the Kruskal-Wallis test, t-test and the Wilcoxon rank sum test (also known as the Man Whitney U-test) were conducted as pairwise tests of independent samples with Bonferroni corrections.

All statistical tests were done with R version 3.4.4 using RStudio version 1.1.456.

4 Results

Detailed results for each measure are presented below. Because of the nature of the data, both means and medians are used to describe the pseudo-longitudinal development trends. To address RQ1, results across proficiency levels are illustrated by boxplots where the group means are also shown. The two learner groups are presented together for each type of task to provide a visual comparison of the differences and similarities between the writer groups and the text types. The numeric values for means, standard deviations, medians, and interquartile ranges for each measure can be found in the tables in Appendix A. To address RQ2, the results of statistical comparisons are summarised in the text, and the post-hoc test results are visualised by tables indicating statistically significant differences between proficiency levels. Section 4.8 provides a summary of the results.

4.1 Mean length of sentence (MLS)

The average length of sentence measured with MLS grows across the proficiency levels, but the increase in length is continuous from the lowest level to the highest only in the formal messages when measured with group means, and in the adolescent learners' argumentative texts when measured with group medians (Figure 1).

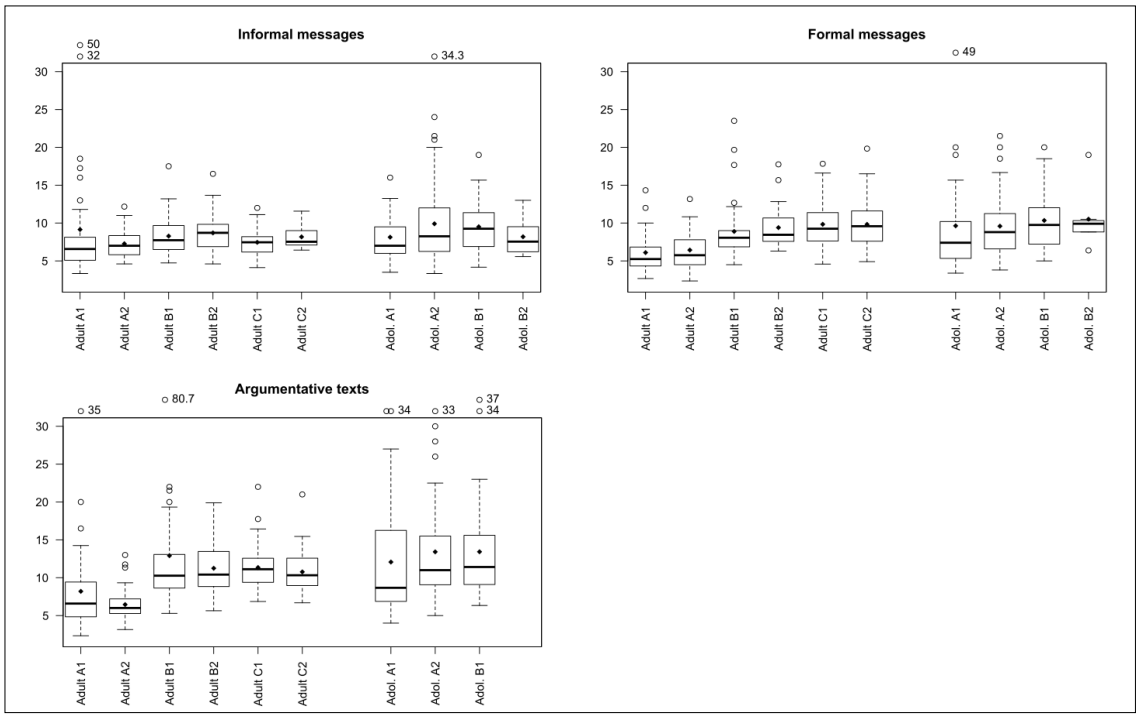

Figure 1. MLS in informal messages, formal messages, and argumentative texts.

According to both one-way ANOVA and the Kruskal-Wallis test, there are statistically significant differences in MLS in the adult learners' formal messages ($F(5,222) = 9.93$, $p < .001$, $R^2_{Adj} = .16$; *Chi squared* = 53.53, $p = < .001$, $df = 5$) and argumentative texts ($F(5,235) = 6.75$, $p < .001$, $R^2_{Adj} = .11$; *Chi squared* = 75.96, $p = < .001$, $df = 5$). According to the Kruskal-Wallis test, there are statistically

significant differences also in adult learners' informal messages (*Chi squared = 12.42, p = .029, df = 5*). The differences between the proficiency levels are not statistically significant in the adolescent learners' data according to either test.

In the adult learners' informal messages, there are no statistically significant differences in the post hoc tests even if the Kruskal-Wallis test result indicates between-group differences in MLS. Table 3 presents the levels between which there is a statistically significant difference according to the post hoc tests.

Table 3. The statistically significant between-level differences in MLS according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1							✓	✓	✓	✓		✓	✓	✓	✓
Adult A2							✓	✓	✓	✓		✓	✓	✓	✓
Adult B1															
Adult B2															
Adult C1															

4.2 Mean length of T-unit (MLTU)

A growing trend of mean length of T-unit (MLTU) is found in the formal messages and in the argumentative texts, where the growth is continuous across all the proficiency levels according to both group means and medians in the formal messages of the adolescent learners, according to group means in the L2 adult learners' formal messages, and according to group medians in the L2 adolescent learners' argumentative texts (Figure 2).

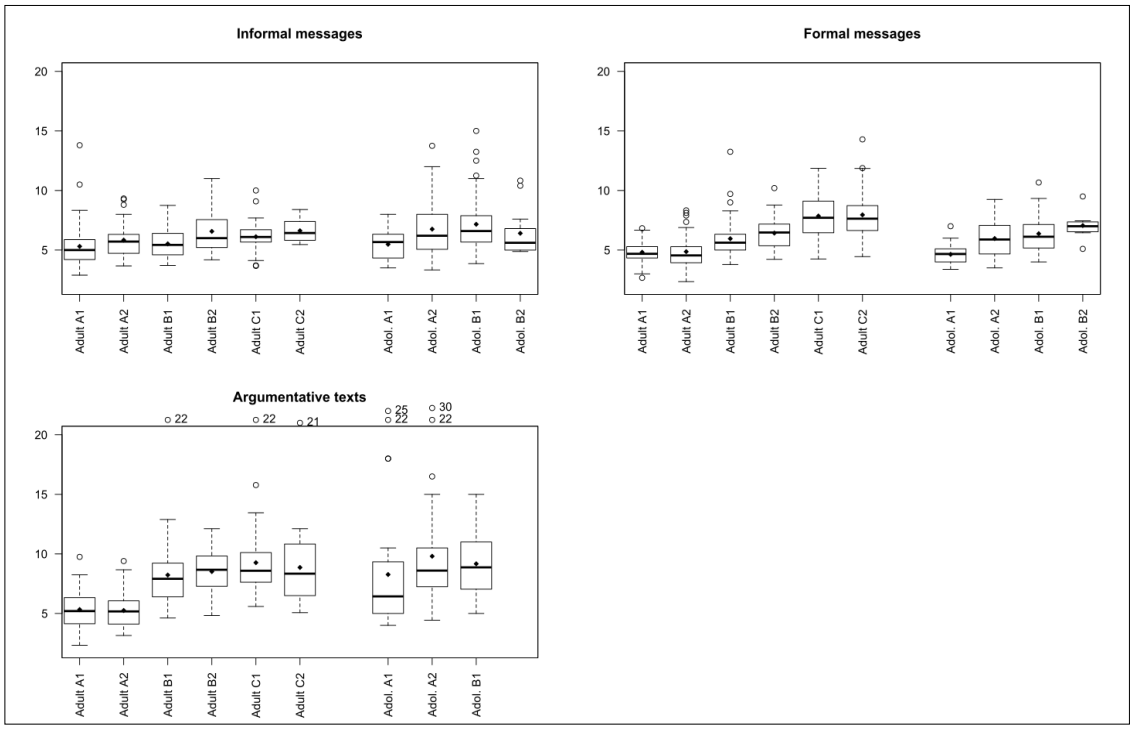

Figure 2. MLTU in informal messages, formal messages, and argumentative texts.

There are statistically significant differences between proficiency levels in MLTU according to both ANOVA and the Kruskal-Wallis test in all the task types in the adult learner data: informal messages (*F(5,192) = 3.73, p = .003, R^2_{Adj} = .06*; *Chi squared = 25.41, p < .001, df = 5*), formal messages (*F(5,222) = 23.95, p < .001, R^2_{Adj} = .34*; *Chi squared = 90.96, p < .001, df = 5*), and argumentative texts (*F(5,235) = 23.91, p < .001, R^2_{Adj} = .32*; *Chi squared = 108.53, p < .001, df = 5*). In the adolescent learner data there are statistically significant between-level differences according to both tests in the informal (*F(3,176) = 3.81, p < .001, R^2_{Adj} = .04*; *Chi squared = 12.48, p = .006, df = 3*) and formal (*F(3,116) = 23.29, p < .001, R^2_{Adj} = .23*; *Chi squared = 33.65, p < .001, df = 3*) messages. There are also statistically significant differences between proficiency levels in the adolescent learners' argumentative texts according to the Kruskal-Wallis test result (*Chi squared = 10.26, p = .006, df = 2)*. The between-level differences which are statistically significant according to the post hoc tests are shown in Table 4.

Table 4. The statistically significant between-level differences in MLTU according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			✓$_U$		✓		✓	✓	✓	✓		✓	✓	✓	✓
Adult A2							✓$_U$	✓	✓	✓		✓	✓	✓	✓
Adult B1			✓$_t$		✓				✓	✓					
Adult B2									✓	✓					
Adult C1															
Adol. A1	✓$_t$	✓				✓	✓	✓			✓$_U$	✓$_U$			
Adol. A2															
Adol. B1															

4.3 Mean length of clause (MLC)

There is growth in the average length of clauses measured with MLC in all three task types (Figure 3). In the informal messages, group means and medians both become higher from the lowest to the highest proficiency level in the adult learners' texts. In the formal messages, the same pattern can be found in the adult learner data (group means) and in the adolescent learner data (group means and medians). In the argumentative texts, both the group means and the medians grow in the adolescent learner texts.

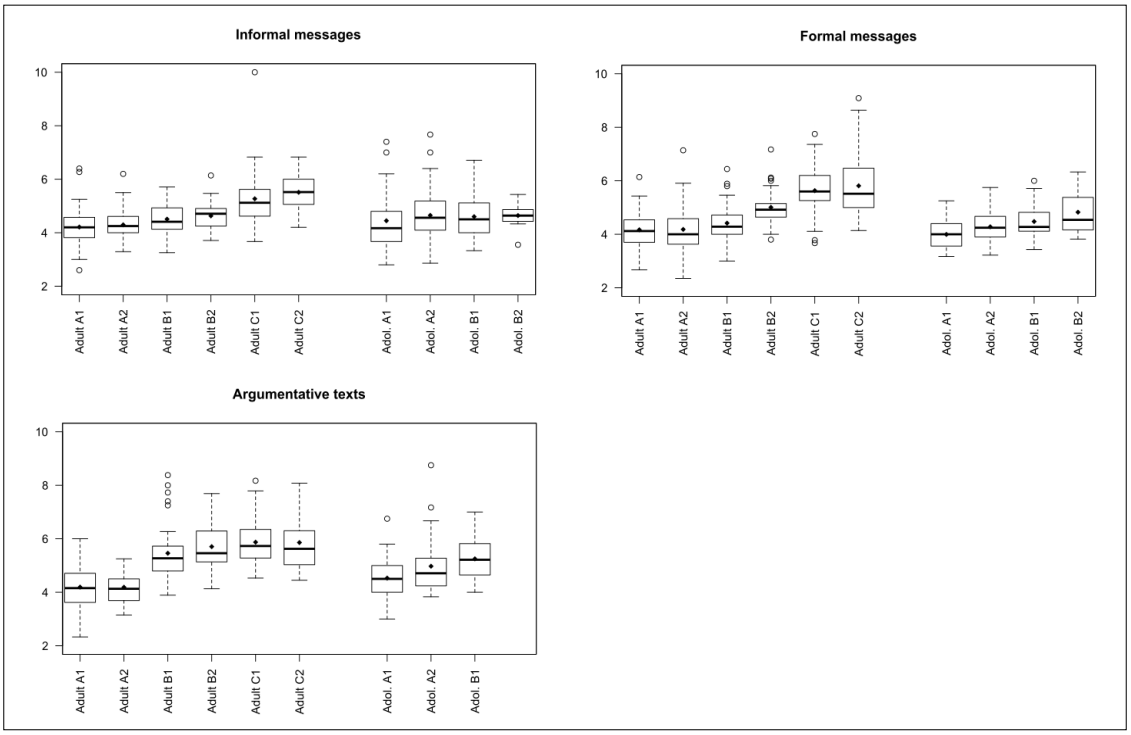

Figure 3. MLC in informal messages, formal messages, and argumentative texts.

There are statistically significant differences according to both ANOVA and the Kruskal-Wallis test between proficiency levels in MLC in the L2 adult learner informal messages ($F(5,192) = 11.95$, $p < .001$, $R^2_{Adj} = .22$; $Chi\ squared = 46.95$, $p < .001$, $df = 5$), formal messages ($F(5,222) = 26.52$, $p < .001$, $R^2_{Adj} = .36$; $Chi\ squared = 95.97$, $p < .001$, $df = 5$), and argumentative texts ($F(5,235) = 35.48$, $p < .001$, $R^2_{Adj} = .42$; $Chi\ squared = 119.07$, $p < .001$, $df = 5$), and also in the L2 adolescent learner formal messages ($F(3,116) = 6.11$, $p < .001$, $R^2_{Adj} = .11$; $Chi\ squared = 14.66$, $p = .002$, $df = 3$) and argumentative texts ($F(2,108) = 6.24$, $p = .003$, $R^2_{Adj} = .09$; $Chi\ squared = 13.22$, $p = .001$, $df = 2$). The post hoc test results are summarised in Table 5.

Table 5. The statistically significant between-level differences in MLC according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			✓$_U$	✓	✓			✓	✓	✓		✓	✓	✓	✓
Adult A2				✓	✓			✓	✓	✓		✓	✓	✓	✓
Adult B1				✓$_U$	✓			✓	✓	✓					
Adult B2					✓				✓	✓					
Adult C1															
Adol. A1							✓					✓			
Adol. A2															
Adol. B1															

4.4 Mean number of T-units per sentence (TU/S)

The patterns of change are mixed for TU/S (Figure 4). The group means and medians grow smaller after the lowest proficiency level, but the change continues from the lowest proficiency level to the highest only in the L2 adolescent learners' informal messages (group means) and formal messages (group medians).

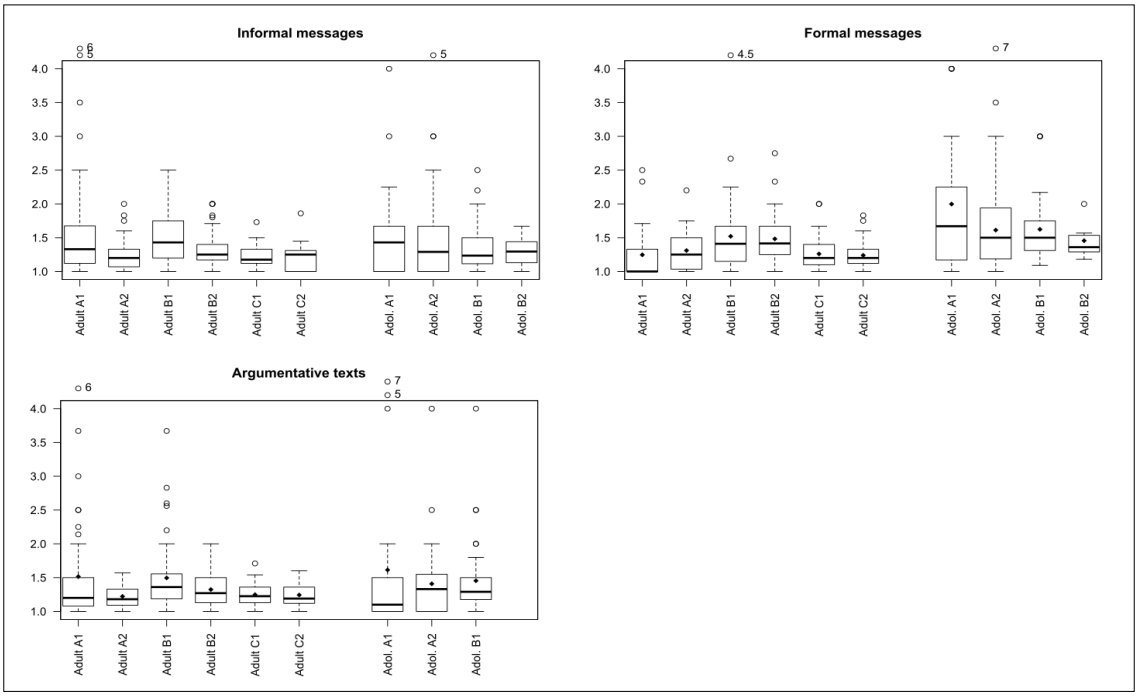

Figure 4. TU/S in informal messages, formal messages, and argumentative texts.

In the adult learners' texts, there are statistically significant differences between proficiency levels in TU/S according to both ANOVA and the Kruskal-Wallis test in the informal ($F(5,192) = 3.55$, $p = .004$, $R^2_{Adj} = .06$; $Chi\ squared = 16.25$, $p < .001$, $df = 5$) and formal ($F(5,222) = 4.62$, $p < .001$, $R^2_{Adj} = .07$; $Chi\ squared = 26.26$, $p < .001$, $df = 5$) messages and according to ANOVA in the argumentative texts ($F(5,235) = 3.24$, $p = .008$, $R^2_{Adj} = .04$). No statistically significant differences between proficiency levels were found in the adolescent learners' texts. The statistically significant between-level differences indicated by the post hoc tests are summarised in Table 6.

Table 6. The statistically significant between-level differences in TU/S according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1								✓$_U$							
Adult A2		✓										✓$_t$			
Adult B1				✓											
Adult B2									✓	✓					
Adult C1															

4.5 Mean number of clauses per sentence (C/S)

The patterns for C/S are mixed (Figure 5). The changes between proficiency levels are not linear, but there are more clauses per sentence on the highest proficiency level compared to the lowest according to group medians in the informal messages and in the argumentative texts in both the adult and adolescent learner data, and according to the group means in the formal messages written by the adult learners. In the adolescent learners' formal messages, the group mean is lower on level B2 than on level A1, and the group median is the same on levels A1 and B2.

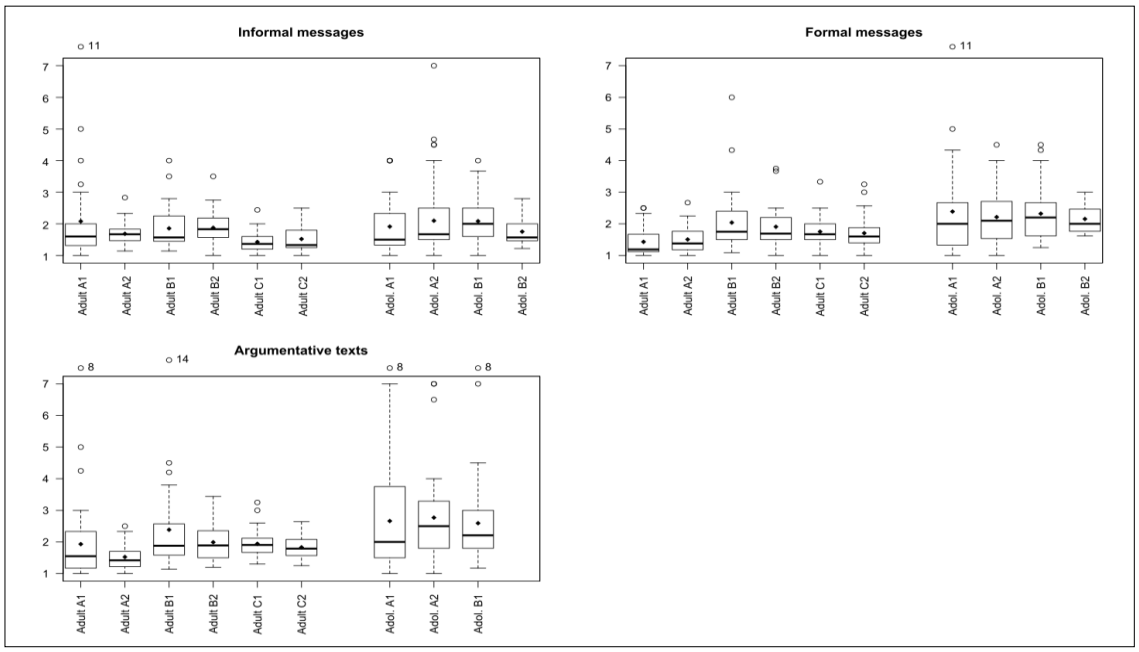

Figure 5. C/S in informal messages, formal messages, and argumentative texts.

There are statistically significant differences in C/S between proficiency levels in L2 adult learner informal messages ($F(5,192) = 2.33$, $p = .044$, $R^2_{Adj} = .03$; *Chi squared* = 20.32, $p < .001$, $df = 5$), formal messages ($F(5,222) = 5.03$, $p < .001$, $R^2_{Adj} = .08$; *Chi squared* = 25.92, $p < .001$, $df = 5$), and argumentative texts ($F(5,235) = 2.74$, $p = .020$, $R^2_{Adj} = .04$; *Chi squared* = 27.51, $p < .001$, $df = 5$) according to both ANOVA and the Kruskal-Wallis test. There are no statistically significant differences between proficiency levels in the L2 adolescent learner texts. The post hoc test results vary according to task type (Table 7).

Table 7. The statistically significant between-level differences in C/S according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1							✓	✓	✓$_U$						
Adult A2				✓			✓	✓$_U$				✓$_U$	✓	✓	✓
Adult B1				✓											
Adult B2				✓											
Adult C1															

4.6 Mean number of clauses per T-unit (C/TU)

For mean number of clauses per T-unit (C/TU), the patterns are varied, and both the group mean and median indicate growth in the number of clauses between the lowest and highest proficiency levels only in the adult learners' formal messages and argumentative texts and in the adolescent learners' formal messages, where the group means grow continuously from one proficiency level to the next (Figure 6).

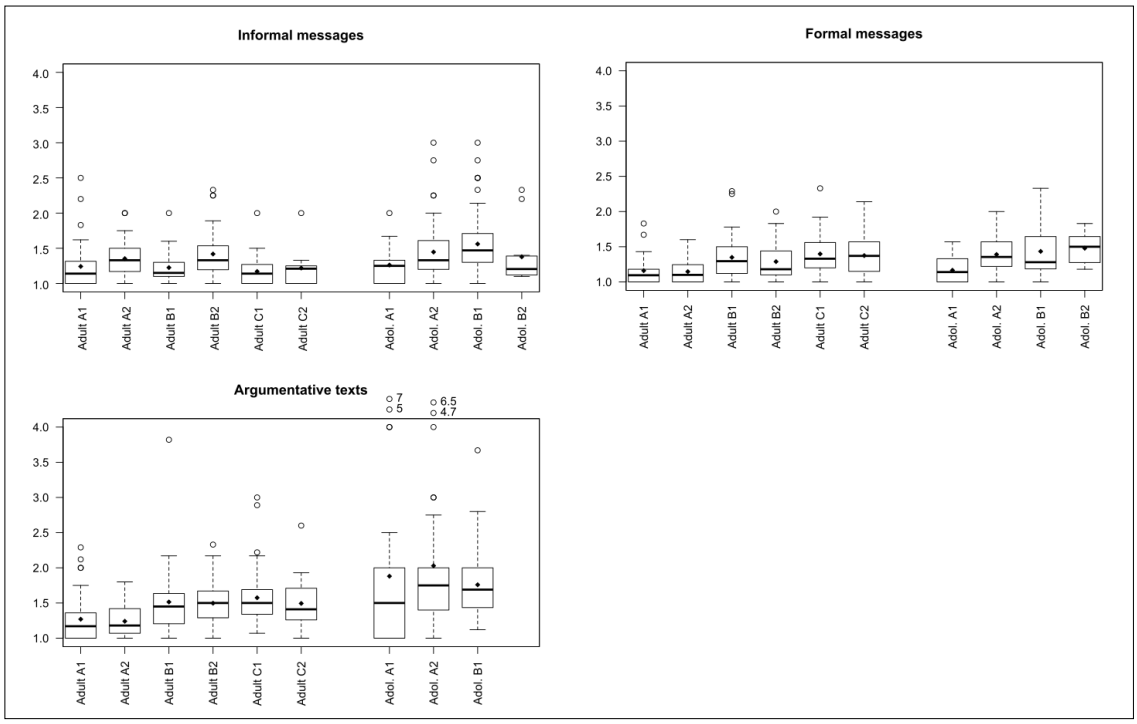

Figure 6. C/TU in informal messages, formal messages, and argumentative texts.

The differences between proficiency levels in C/TU are statistically significant according to both ANOVA and the Kruskal-Wallis test in the informal messages of both the adult ($F(5,192) = 3.98$, $p = .002$, $R^2_{Adj} = .07$; $Chi\ squared = 26.04$, $p < .001$, $df = 5$) and the adolescent learners ($F(3,176) = 3.78$, $p = .012$, $R^2_{Adj} = .04$; $Chi\ squared = 13.55$, $p = .004$, $df = 3$), in the formal messages of both the adult ($F(5,222) = 5.66$, $p < .001$, $R^2_{Adj} = .09$; $Chi\ squared = 34.13$, $p < .001$, $df = 5$) and adolescent learners ($F(5,235) = 6.69$, $p < .001$, $R^2_{Adj} = .11$; $Chi\ squared = 44.73$, $p < .001$, $df = 5$) and in the adult learners' argumentative texts ($F(5,235) = 6.69$, $p < .001$, $R^2_{Adj} = .11$; $Chi\ squared = 44.73$, $p < .001$, $df = 5$). Table 8 shows the between-level differences that are statistically significant according to the post hoc tests.

Table 8. The statistically significant between-level differences in C/TU according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			✓$_U$				✓$_U$		✓	✓		✓$_U$	✓	✓	✓$_U$
Adult A2							✓		✓	✓		✓	✓	✓	✓

Adult B1			✓												
Adult B2				✓											
Adult C1															
Adol. A1	✓$_t$	✓				✓	✓	✓$_U$							
Adol. A2															
Adol. B1															

4.7 Mean number of dependent clauses per clause (DC/C)

The group means and medians of DC/C show that there are more dependent clauses per clause on the highest proficiency level than on the lowest in all task types, with the exception of the group medians of the adolescent learners' informal messages. In the adolescent learners' formal messages, the growth in group means is continuous from level A1 to B2 (Figure 7).

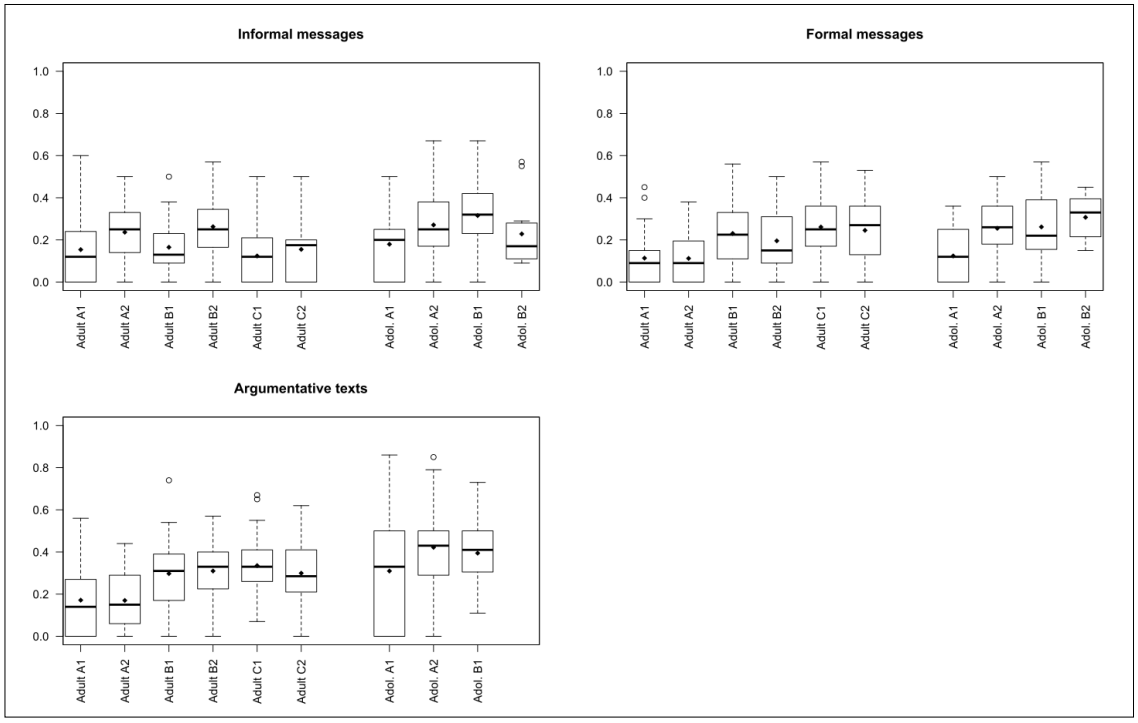

Figure 7. DC/C in informal messages, formal messages, and argumentative texts.

In DC/C, there are statistically significant differences between proficiency levels according to both ANOVA and the Kruskal-Wallis test in the informal messages of both the adult ($F(5,192) = 5.23$, $p < .001$, $R^2_{Adj} = .10$; *Chi squared* $= 25.83$, $p < .001$, $df = 5$) and adolescent learners ($F(3,176) = 4.62$, $p = .004$, $R^2_{Adj} = .06$; *Chi squared* $= 13.62$, $p = .004$, $df = 3$), in the formal messages of both the adult ($F(5,222) = 7.47$, $p < .001$, $R^2_{Adj} = .12$; *Chi squared* $= 33.94$, $p < .001$, $df = 5$) and adolescent learners ($F(2,116) = 7.61$, $p < .001$, $R^2_{Adj} = .14$; *Chi squared* $= 19.93$, $p < .001$, $df = 3$), and in the adult learners' argumentative texts ($F(5,235) = 10.93$, $p < .001$, $R^2_{Adj} = .17$; *Chi squared* $= 44.87$, $p < .001$, $df = 5$). Table 9 presents the levels between which there is a statistically significant difference according to the post hoc tests.

Table 9. The statistically significant between-level differences in DC/C according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					Formal messages					Argumentative texts				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			✓				✓		✓	✓		✓	✓	✓	✓
Adult A2				✓			✓		✓	✓		✓	✓	✓	✓
Adult B1			✓												
Adult B2				✓											
Adult C1															
Adol. A1		✓				✓	✓	✓							
Adol. A2															
Adol. B1															

4.8 Summary of results

The measures yield different results for adult and adolescent learners and also for the different task types. In the adult learner data, there are statistically significant differences between proficiency levels in all the seven measures, while in the L2 adolescent learner data, only four of the measures, i.e. MLTU, MLC, C/TU, and DC/C, show statistically significant differences between proficiency levels. In the adult learner data, there are no statistically significant between-level differences in MLS in the informal messages. In the adolescent learner data, MLTU is the only measure showing statistically significant between-level differences in all three task types.

Developmental trends in group means may differ from those in group medians. There is growth in group means between the lowest and highest proficiency levels in MLTU, MLC and DC/C in all three task types in both the adult and adolescent learner data. The same kind of growth is found in group medians in MLS and MLC. When comparing either group means or group medians, at least one of them is higher on the highest proficiency level than on the lowest also in C/TU, and is lower in TU/S in all task types for both the adult and adolescent learners. None of the measures indicating growth reach the highest values at the highest proficiency level in all three task types, and the change in TU/S is continuous from one proficiency level to the next only in the group means of the adolescent learners' informal messages, where the between-level differences are not statistically significant.

Most measures show statistically significant differences between levels A1 and C2 in the adult learner data, although the differences between levels A1 and C2 are not statistically significant in TU/S and C/S in any of the task types and in MLS, C/TU and DC/C they are statistically significant only in the formal messages and argumentative texts. In the adolescent learner data, there are statistically significant differences between level A1 and the highest proficiency level in MLTU, C/TU and DC/C in the formal messages, and in MLTU and MLC in the argumentative texts.

There is variation in the measures' ability to gauge differences between adjacent proficiency levels. In the adult learner data, none of the measures show

statistically significant differences between levels A1 and A2 or between levels C1 and C2. In the adolescent learner data, all the statistically significant differences are between level A1 and the levels above it.

The effect sizes for all measures showing statistically significant between-level differences in one-way ANOVA are at least small (> .01) when measured with adjusted R squared. The effect size is large (> .14) in the L2 adult learner data for MLC in all task types, for MLTU and DC/C in the formal messages and argumentative texts, and for MLS in the formal messages. In the L2 adolescent learner data, the effect size is large for MLTU and DC/C in the formal messages. The effect size is medium (>. 06) in the L2 adult learner data for C/TU in all task types, for TU/S in the informal and formal messages, for MLS in the argumentative texts, for MLTU and DC/C in the informal messages, and for C/S in the formal messages. In the L2 adolescent learner data, effect size is medium for MLC in the formal messages and argumentative texts and for C/TU in the formal messages. For the measures showing no statistically significant differences between proficiency levels, the effect size is small for TU/S in the L2 adolescent learner formal messages. For the remaining measures, the effect size is less than small.

5 Discussion and conclusions

In the present study, syntactic complexity and seven quantitative measures were studied in relation to the CEFR proficiency levels in cross-sectional data of learner Finnish. Both measures for overall complexity (MLS and MLTU), the measure for sub-clausal complexity (MLC), and the two subordination measures (C/TU and DC/C) grow from the lowest proficiency level to the highest, and the coordination-based measure (TU/S) diminishes from the lowest proficiency level to the highest even if the changes are not linear. The growing trend is in line with earlier research findings, such as those included in the synthesis of Wolfe-Quintero et al. (1998). The general reduction in coordination from level A1 to the highest proficiency level is in line with the notion of clausal coordination being more typical of lower proficiency levels (e.g. Norris & Ortega, 2009).

There are statistically significant differences between proficiency levels only for some of the measures and only between some proficiency levels. Indeed, the measures often have overlapping values when proficiency levels are compared. In the adult learner data, there are differences between the beginners and advanced learners in most measures, but the intermediate learners' results overlap with the beginners' or advanced learners' results, or both, in all the measures. In the adolescent learner data, there are overlapping results in all the measures.

Regarding the first research question, *How does syntactic complexity in written learner Finnish develop across CEFR proficiency levels when measured quantitatively*, the results suggest that the development is different in the adult learner data and in the adolescent learner data. The results show that there are statistically significant between-level differences in all the measures in the adult learner data, but in only four measures in the adolescent learner data. There is also more within-group variation in many of the measures in the adolescent learner data than in the adult learner data. These findings suggest that adult and adolescent

L2 Finnish learners use some syntactic features differently in their writing. The measures also indicate different patterns of development in the two age groups: in the adult learner data, the statistically significant differences are typically found when levels A1 and A2 are compared to the higher proficiency levels, whereas in the adolescent learner data, the differences are between level A1 and the other proficiency levels. Similar differences between adult and adolescent learners have been found in the use of existential sentences (Kajander, 2013), indirect references (Seilonen, 2013), and transitive constructions (Reiman, 2014) in the same data.

The answer to the second research question, *How well do the quantitative measures used in this study differentiate CEFR proficiency levels in written learner Finnish*, is mixed. While most of the measures do differentiate the lowest proficiency levels from the highest, most of them do not differentiate the intermediate learner levels from other levels or differentiate between adjacent proficiency levels. In this regard, mean length of clause (MLC) seems the best measure of syntactic complexity for adult L2 learner Finnish, as it develops quite linearly and it is also able to differentiate the intermediate levels from the levels both below and above. For adolescent L2 learner Finnish, mean length of T-unit (MLTU) seems the best measure, as it is able to differentiate level A1 from most of the levels above it in all task types, even if the increase in MLTU is not linear in all task types. The results also suggest that the measures are sensitive to task type, which is in line with findings from other studies (see e.g. Michel, 2017). In the present study, there are statistically significant differences between proficiency levels in all seven measures in the adult learners' formal messages and argumentative texts, but only in six measures in the informal messages. In the adolescent learner data, only four measures show statistically significant differences between proficiency levels; four of them in the formal messages, three in the informal messages, and two in the argumentative texts. The formal messages also seem to show statistically significant differences between more proficiency levels than do the other two task types. Differences between the task types in the CEFLING corpus have also been found by Seilonen (2013), Kajander (2013), and Reiman (2014).

There are a number of limitations to this study. First, the data were segmented into the production units by one annotator only. As learner language contains deviations from the norms, annotating learner language always involves some level of interpretation of the intended forms (e.g. Brunni, Lehto, Jantunen, & Airaksinen, 2015; Granger, 2002). Segmenting the data used in this study into clauses, sentences, and T-units is no exception (Martin, 2013; Mylläri, 2020). Using more annotators and negotiating the segmentation could result in different numbers of clauses, sentences, and T-units in some texts, and this could affect the measures.

Second, the texts used in this study are relatively short. This can partly be explained by typological features of Finnish, such as the lack of articles and the limited use of prepositions, which affect the word count. On average, the length of the texts written by the adolescent learners corresponds to that of the texts written by their L1 counterparts in the CEFLING project. The adult learners' texts are generally longer than those by the adolescent learners, and the adult learners' argumentative texts reach the length of 100 words or more at level B1. This corresponds to the length of the 100-word random samples used in the studies of Spoelman and Verspoor (2010) and Tilma (2014), who also used

shorter samples at the beginning, when the learner texts did not reach 100 words.

Third, there would be grounds for criticising the statistical analysis of the data. There is a varying amount of individual variation in the data and there are outliers in many of the groups that were compared. Not excluding the outliers from the statistical analysis could have had an impact on the parametric tests used in the study. The effect was partly controlled by using both parametric and non-parametric tests. Also, the results for adolescent learner proficiency level B2 should be interpreted with caution since there is only a limited number of texts on that level. Therefore the statistical significance or insignificance of the results should not be interpreted as straightforward evidence of the measures' general ability or inability to gauge differences between proficiency levels.

The present study suggests some interesting topics for future research. A closer analysis of the differences between adult and adolescent L2 learners, as well as of the differences between the texts written by the adolescent L2 learners and the corresponding L1 writers, could be worthwhile. Another topic for future research could be the correlation between the measures (cf. Lu, 2017), as they may prove more powerful indicators of proficiency together than individually. Also, the present study focused on syntactic complexity in cross-sectional data and in relation to proficiency. The results therefore cannot be interpreted as reflecting the measures' value as indicators of development, which should be studied using longitudinal data.

The results of this study support calls for new ways of exploring complexity, especially in morphologically rich languages. For learner Finnish, Reiman (2011b) has argued for a more qualitative approach, and Tilma (2014) has used morphemes instead of words in length-based quantitative measures. Although measures of complexity cannot be validated by simply showing increase across time or proficiency (Bulté and Housen, 2012), and increasing complexity does not necessarily mean increasing proficiency, as pointed out by Ortega (2003) and Pallotti (2009), there is a need for new means to gauge complexity across proficiency levels if syntactic complexity is going to be used to measure learner language proficiency.

Endnote

[1] The parsing pipeline and the clause splitting feature are available under an open licence at http://turkunlp.github.io/Finnish-dep-parser/. A version available on July 29, 2018 was used.

References

Alanen, R., Huhta, A., & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 21–56). EUROSLA Monographs series 1.

Alisaari, J. (2016). *Songs and poems in the second language classroom. The hidden potential of singing for developing writing fluency.* Annales Universitaris Turkuensis, Series B Humaniora 426. Turku: University of Turku.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *Tesol Quarterly, 26(2)*, 390–395.

Bernardini, P., & Granfeldt, J. (2019). On crosslinguistic variation and measures of linguistic complexity in learner texts: Italian, French and English. *International Journal of Applied Linguistics, Special issue*, 211–232.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly, 45(1)*, 5–35.

Brunni, S., Lehto, L., Jantunen, J., & Airaksinen, V. (2015). How to annotate morphologically rich learner language. Principles, problems and solutions. *Bergen Language and Linguistics Studies 6*, 133–152.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins Publishing Company.

Council of Europe. (2001). *Common European framework of reference for: learning, teaching, assessment*. Retrieved from https://rm.coe.int/1680459f97

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language.* Oxford: Oxford University Press.

Foster, P., Tonkyn A., & Wigglesworth G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics 21(3)*, 354–375.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: Benjamins.

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction*. (2nd rev. ed.). Berlin: De Gruyter Mouton.

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. In L. Roberts, I. Vedder, & J.H. Hulstijn (Eds.), *Eurosla Yearbook 14* (pp. 1–30). Amsterdam: John Benjamins.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., & Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation 48*, 493–531.

Housen, A., De Clercq, B., Kuiken, F., & Vedder I. (2019). Multiple approaches to complexity in second language research, *Second Language Research, 35(1)*, 3–21.

Housen, A., Kuiken, V., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA.* Amsterdam: John Benjamins Publishing Company.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study – Validating the rating process across tasks, scales and languages. *Language Testing, 31(3)*, 307–328.

Jantunen, J., & Pirkola, S. (2015). Oppijansuomen sähköiset tutkimusaineistot. Nykytilanne [Electronic corpora of learner Finnish. Current situation]. *Virittäjä, 119(1),* 88–103 .

Kajander, M. (2013). *Suomen eksistentiaalilause toisen kielen oppimisen polulla* [Paths of learning Finnish existential sentences]. Jyväskylä studies in humanities 220. Jyväskylä: University of Jyväskylä.

Kalliokoski, J. (2006). Virke, dialogisuus ja argumentaatio: irralliset sivulauseet ja toisella kielellä kirjoittaminen [Sentence, dialogue and argumentation: stand-alone subordinate clauses and second language writing]. In T. Nordlund, T. Onikki-Rantajääskö, T. Suutari, & H. Forsberg (Eds.), *Kohtauspaikkana kieli: näkökulmia persoonaan, muutoksiin ja valintoihin* [Language as a meeting place: Perspectives on person, changes and choices] (pp. 212–231). Helsinki: Suomalaisen Kirjallisuuden Seura.

Kuiken, F. & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish, *International Journal of Applied Linguistics, Special issue,* 192–210.

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

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

Larson-Hall J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

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

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

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing, 34(4)*, 493–511.

Martin, M. (2013). Sentences and clauses as complexity measures in second language writing: a segmentation experiment. In M. Järventausta, & M. Pantermöller (Eds.), *Finnische Sprache, Literatur und Kultur im deutschsprachigen Raum – Suomen kieli, kirjallisuus ja kulttuuri saksankielisellä alueella* (pp. 185–198). Greifswald: Veröffentlichungen der Societas Uralo-Altaica.

Martin, M., Mustonen, S., Reiman, N., & Seilonen, M. (2010). On becoming an independent user. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 57–79). EUROSLA Monographs series 1.

Michel, M. C. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen, & S. Masatoshi (Eds.), R*outledge handbook of instructed second language acquisition* (pp. 50–68). New York: Routledge.

Mylläri, T. (2020). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association, 4(1),* 13–23.

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

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

Pallotti, G. (2009). CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics, 30(4)*, 590–601.

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

Reiman, N. (2011a). Transitiivikonstruktio ikkunana syntaksin kehitykseen: infiniittiset rakenteet ja passiivi taidon indikaattoreina S2-oppijoiden teksteissä [The transitive construction as a window into syntax development: Infinite structures and passive as indicators of proficiency in F2 students' texts]. In E. Lehtinen, S. Aaltonen, M. Koskela, E. Nevasaari, & M. Skog-Södersved (Eds.), *AFinLAe 3* (pp. 142–157). Retrieved from http://ojs.tsv.fi/index.php/afinla/issue/view/694

Reiman, N. (2011b): Two faces of complexity: structural measures and diversity of constructions. *Nordand, 6(2)*, 9–23.

Reiman, N. (2014). Yläkoulun S2-oppilaiden transitiivi-ilmausten käyttö Eurooppalaisen viitekehyksen taitotasoilla [Lower secondary school L2 Finnish students' use of transitive expressions at the CEFR levels]. *Lähivõrdlusi. Lähivertailuja 24*, 183–220.

Seilonen, M. (2013). *Epäsuora henkilöön viittaaminen oppijansuomessa* [Indirect references in Finnish learner language] Jyväskylä Studies in Humanities 197. Jyväskylä: University of Jyväskylä.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31(4)*, 532–553.

Tilma, C. (2014). *The dynamics of foreign versus second language development in Finnish writing*. Jyväskylä studies in humanities 233. Jyväskylä: University of Jyväskylä.

Verspoor, M., Lowie, W., Chan, H., & Vahtrick, L. (2017). Linguistic complexity in second language development: variability and variation at advanced stages. *Recherches en didactique des langues et des cultures, 14(1)*, 1–27

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity. Technical report No. 1.* Honolulu: Second Language Teaching and Curriculum Center.

Appendices

Appendix A.

Table A1. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, mean length of sentence (MLS).

MLS	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	9.14	8.55	6.57	3.04	22	6.10	2.86	5.24	2.39	50	8.20	5.44	6.59	4.54
Adult A2	39	7.27	1.93	7.00	2.52	27	6.43	2.68	5.75	3.29	37	6.45	2.20	6.00	1.93
Adult B1	41	8.28	2.66	7.73	3.17	42	8.89	3.78	8.06	2.13	43	12.91	11.30	10.27	4.44
Adult B2	39	8.70	2.58	8.71	2.95	34	9.40	2.59	8.45	2.98	35	11.25	3.27	10.41	4.63
Adult C1	26	7.45	1.96	7.46	1.90	45	9.84	2.75	9.25	3.76	46	11.35	2.78	11.12	3.11
Adult C2	14	8.17	1.58	7.53	1.65	58	9.83	2.79	9.58	3.82	30	10.77	3.02	10.32	3.40
Adol. A1	25	8.13	3.37	7.00	3.50	33	9.63	8.22	7.40	4.87	32	12.07	8.46	8.67	8.44
Adol. A2	79	9.90	5.40	8.25	5.75	40	9.59	4.19	8.80	4.50	39	13.41	6.73	11.00	6.44
Adol. B1	64	9.51	3.17	9.25	4.36	40	10.34	3.78	9.75	4.53	40	13.42	6.71	11.42	6.15
Adol. B2	12	8.19	2.56	7.55	2.83	7	10.51	3.99	9.91	1.50	-	-	-	-	-

Table A2. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, mean length of T-unit (MLTU).

MLTU	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	5.31	2.07	5.00	1.67	22	4.82	1.14	4.69	0.93	50	5.33	1.61	5.20	2.11
Adult A2	39	5.84	1.49	5.71	1.59	27	4.86	1.59	4.55	1.35	37	5.25	1.47	5.17	1.95
Adult B1	41	5.52	1.15	5.43	1.80	42	5.95	1.69	5.62	1.29	43	8.23	2.87	7.91	2.83
Adult B2	39	6.57	1.67	6.00	2.35	34	6.43	1.37	6.48	1.78	35	8.52	1.93	8.67	2.55
Adult C1	26	6.12	1.43	6.10	1.01	45	7.86	1.81	7.71	2.66	46	9.26	2.86	8.59	2.41
Adult C2	14	6.62	0.97	6.43	1.54	58	7.95	2.00	7.64	2.09	30	8.86	3.16	8.34	4.32
Adol. A1	25	5.48	1.19	5.67	2.00	33	4.64	0.81	4.67	1.10	32	8.27	5.26	6.44	2.45
Adol. A2	79	6.76	2.22	6.20	2.94	40	5.97	1.43	5.89	2.37	39	9.80	4.75	8.60	2.94
Adol. B1	64	7.16	2.31	6.60	2.11	40	6.37	1.59	6.12	1.92	40	9.17	2.54	8.87	2.75
Adol. B2	12	6.41	2.11	5.61	1.35	7	7.06	1.33	7.00	0.83	-	-	-	-	-

Table A3. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, mean length of clause (MLC).

MLC	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	4.21	0.77	4.20	0.76	22	4.17	0.80	4.12	0.82	50	4.19	0.76	4.16	1.08
Adult A2	39	4.30	0.62	4.25	0.61	27	4.18	0.98	4.00	0.96	37	4.19	0.61	4.13	0.81
Adult B1	41	4.51	0.62	4.41	0.80	42	4.42	0.71	4.29	0.72	43	5.46	1.03	5.27	0.93
Adult B2	39	4.63	0.50	4.71	0.66	34	5.01	0.68	4.92	0.48	35	5.71	0.90	5.46	1.16
Adult C1	26	5.27	1.22	5.12	0.96	45	5.63	0.84	5.60	0.94	46	5.87	0.86	5.73	1.04
Adult C2	14	5.51	0.72	5.52	0.86	58	5.81	1.09	5.52	1.42	30	5.86	1.04	5.63	1.26
Adol. A2	79	4.65	0.85	4.56	1.09	40	4.28	0.59	4.25	0.77	39	4.97	1.01	4.71	1.04
Adol. B1	64	4.60	0.78	4.50	1.09	40	4.48	0.56	4.28	0.66	40	5.25	0.81	5.22	1.15
Adol. B2	12	4.64	0.49	4.64	0.36	7	4.83	0.91	4.54	1.22	-	-	-	-	-

Table A4. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, T-units per sentence (TU/S).

TU/S	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	1.67	1.06	1.33	0.56	22	1.25	0.43	1.00	0.30	50	1.51	0.86	1.20	0.40
Adult A2	39	1.26	0.25	1.20	0.26	27	1.31	0.31	1.25	0.47	37	1.22	0.17	1.18	0.24
Adult B1	41	1.50	0.37	1.43	0.55	42	1.52	0.61	1.41	0.50	43	1.50	0.54	1.36	0.37
Adult B2	39	1.33	0.28	1.25	0.23	34	1.48	0.35	1.42	0.40	35	1.32	0.26	1.27	0.37
Adult C1	26	1.22	0.19	1.18	0.21	45	1.26	0.25	1.20	0.30	46	1.25	0.17	1.23	0.23
Adult C2	14	1.25	0.23	1.25	0.28	58	1.24	0.19	1.20	0.20	30	1.24	0.16	1.19	0.24
Adol. A1	25	1.52	0.72	1.43	0.67	33	2.00	1.26	1.67	1.08	32	1.61	1.31	1.10	0.50
Adol. A2	79	1.45	0.61	1.29	0.67	40	1.61	0.61	1.50	0.72	39	1.41	0.55	1.33	0.55
Adol. B1	64	1.35	0.34	1.24	0.38	40	1.62	0.47	1.50	0.43	40	1.45	0.55	1.29	0.32
Adol. B2	12	1.29	0.22	1.30	0.30	7	1.46	0.28	1.36	0.25	-	-	-	-	-

Table A5. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, clauses per sentence (C/S).

C/S	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	2.08	1.69	1.60	0.69	22	1.43	0.49	1.19	0.54	50	1.93	1.21	1.55	1.12
Adult A2	39	1.69	0.34	1.67	0.36	27	1.51	0.43	1.38	0.59	37	1.52	0.40	1.42	0.48
Adult B1	41	1.86	0.63	1.57	0.80	42	2.04	0.89	1.75	0.87	43	2.39	1.97	1.88	0.99
Adult B2	39	1.88	0.50	1.83	0.62	34	1.91	0.59	1.69	0.68	35	1.99	0.58	1.89	0.86
Adult C1	26	1.43	0.32	1.37	0.39	45	1.75	0.44	1.67	0.50	46	1.94	0.42	1.91	0.44
Adult C2	14	1.52	0.43	1.33	0.46	58	1.71	0.44	1.60	0.47	30	1.83	0.35	1.79	0.48
Adol. A1	25	1.91	0.94	1.50	1.00	33	2.39	1.87	2.00	1.34	32	2.66	1.86	2.00	2.13
Adol. A2	79	2.10	1.01	1.67	1.00	40	2.21	0.83	2.10	1.14	39	2.77	1.45	2.50	1.49
Adol. B1	64	2.08	0.67	2.00	0.90	40	2.32	0.85	2.20	1.03	40	2.59	1.37	2.21	1.20
Adol. B2	12	1.75	0.46	1.57	0.42	7	2.15	0.53	2.00	0.70	-	-	-	-	-

Table A6. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, clauses per T-unit (C/TU).

C/TU	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	1.24	0.33	1.14	0.32	22	1.16	0.22	1.10	0.18	50	1.27	0.33	1.17	0.35
Adult A2	39	1.35	0.25	1.33	0.33	27	1.15	0.16	1.10	0.25	37	1.24	0.23	1.18	0.35
Adult B1	41	1.23	0.20	1.15	0.20	42	1.35	0.29	1.30	0.37	43	1.51	0.47	1.45	0.43
Adult B2	39	1.42	0.33	1.33	0.34	34	1.29	0.26	1.18	0.33	35	1.50	0.29	1.50	0.38
Adult C1	26	1.17	0.22	1.14	0.26	45	1.40	0.27	1.33	0.36	46	1.57	0.39	1.50	0.34
Adult C2	14	1.22	0.25	1.21	0.23	58	1.37	0.27	1.37	0.40	30	1.49	0.34	1.41	0.44
Adol. A1	25	1.26	0.25	1.25	0.33	33	1.17	0.18	1.14	0.33	32	1.88	1.34	1.50	1.00
Adol. A2	79	1.45	0.38	1.33	0.41	40	1.39	0.26	1.36	0.35	39	2.03	1.06	1.75	0.60
Adol. B1	64	1.56	0.44	1.47	0.41	40	1.43	0.37	1.28	0.44	40	1.76	0.49	1.69	0.56
Adol. B2	12	1.38	0.43	1.21	0.27	7	1.48	0.24	1.50	0.37	-	-	-	-	-

Table A7. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, dependent clauses per clause (DC/C).

DC/C	Informal message					Formal message					Argumentative text				
	n	**M**	**SD**	**Mdn**	**IQR**	**n**	**M**	**SD**	**Mdn**	**IQR**	**n**	**M**	**SD**	**Mdn**	**IQR**
Adult A1	39	0.15	0.16	0.12	0.24	22	0.11	0.13	0.09	0.15	50	0.17	0.17	0.14	0.27
Adult A2	39	0.24	0.13	0.25	0.19	27	0.11	0.11	0.09	0.20	37	0.17	0.14	0.15	0.23
Adult B1	41	0.17	0.12	0.13	0.14	42	0.23	0.14	0.23	0.21	43	0.30	0.15	0.31	0.22
Adult B2	39	0.26	0.14	0.25	0.18	34	0.20	0.14	0.15	0.21	35	0.31	0.13	0.33	0.18
Adult C1	26	0.12	0.12	0.12	0.20	45	0.26	0.13	0.25	0.19	46	0.34	0.13	0.33	0.15
Adult C2	14	0.16	0.14	0.18	0.18	58	0.25	0.14	0.27	0.22	30	0.30	0.14	0.29	0.20
Adol. A1	25	0.18	0.15	0.20	0.25	33	0.12	0.12	0.12	0.25	32	0.31	0.27	0.33	0.50
Adol. A2	79	0.27	0.16	0.25	0.21	40	0.26	0.14	0.26	0.18	39	0.42	0.20	0.43	0.21
Adol. B1	64	0.32	0.17	0.32	0.19	40	0.26	0.17	0.22	0.22	40	0.39	0.14	0.41	0.19
Adol. B2	12	0.23	0.17	0.17	0.17	7	0.31	0.11	0.33	0.18	-	-	-	-	-

Received September 18, 2019
Revision received April 25, 2020
Accepted June 26, 2020