Title: Lexical characteristics of young L2 English learners' narrative writing at the start of formal instruction.

Author 1: Vanessa De Wilde
Affiliations author 1: Department of Translation, Interpreting and Communication, Ghent University; Artevelde University of Applied Sciences Ghent

Address for correspondence:
Vanessa De Wilde
Groot-Brittanniëlaan 45
9000 Ghent
Belgium
vanessa.dewilde@ugent.be

To be published in Journal of Second Language Writing

Title: Lexical characteristics of young L2 English learners' narrative writing at the start of formal instruction.

Abstract

Studies investigating L2 English receptive and productive vocabulary knowledge in young learners have shown that English can be picked up through exposure outside the classroom. In this study, I looked into lexical characteristics of young learners' writing at the start of formal English lessons in the first year of secondary school (n = 3168). The texts were given a holistic score and several lexical measures were calculated. The results showed large individual differences between learners' writing. Regression analysis was used to investigate which lexical characteristics predicted proficiency scores. The final model explained 50% of the variance. Similar to what was found in previous research investigating young L2 English learners' writing I found that a number of broad predictors impacted the proficiency score. These were lexical diversity, word count, total number of spelling errors and percentage of English words used. Additionally, four fine-grained variables predicted the proficiency score: word frequency, trigram frequency, age of acquisition and imageability. The results show the added value of investigating a wide range of variables to shed light on the lexical factors that might impact writing scores, even in beginner and pre-intermediate level L2 writing.

Keywords: young learners, complexity, accuracy, writing development, vocabulary

1. Introduction

It is well-established that foreign language learning cannot take place without a sufficient amount of input (Bybee & Hopper, 2001; Ellis, 2002). Past research stressed that language learning in the classroom is often insufficient to become a successful language learner and many studies have looked into ways of increasing L2 learners' input through incidental or

contextual language learning, which is learning a language while focusing on another task such as during extensive reading or watching television (Elgort et al., 2018; Hulstijn, 2003).

In the past decade, a number of studies have been conducted which looked into contextual L2 English language learning which is not teacher-initiated. This type of language learning focusses on the role of out-of-school exposure or extramural English (a term coined by Sundqvist in 2009), which refers to different types of activities that can be done in L2 English such as watching television, gaming or using social media. The studies have consistently shown that these activities, which are initiated by the learners themselves and take place in an informal context, can lead to significant language learning gains (see Zhang et al., 2021 for a review of 33 studies on extramural English learning).

When doing a writing task, many aspects of vocabulary knowledge need to be activated at the same time (Schmitt, 2019) but not much research has looked into young L2 English learners' vocabulary use in a productive task. To the best of my knowledge no studies have been conducted which investigate L2 English language learners' vocabulary use in a writing task at the start of formal classroom instruction. Recent studies did show that young learners pick up English vocabulary through out-of-school exposure (e.g., Bollansée et al., 2020; De Wilde et al., 2020a; Puimège and Peters, 2019). These studies have shown that some children can recognize and produce English words in meaning recognition, form recall and meaning recall tests. The studies have also shown large individual differences between young learners. However, there are no studies which have investigated whether and how this vocabulary knowledge can be used in L2 English writing at the start of formal English lessons. This means that, even though recent studies into language learning through out-of-school exposure have investigated vocabulary knowledge, only little is known about this population's vocabulary use. In this study I will try to fill this gap and explore lexical properties of young learners' L2 English narrative writing in a picture-based task. I will also investigate which

lexical characteristics can discriminate between proficiency levels assigned by L2 English teachers.

2. Background

2.1.Learning vocabulary through out-of-school exposure

A number of studies investigating language learning through out-of-school exposure looked into young L2 English learners' vocabulary knowledge. Some studies looked into the impact of out-of-school exposure on learners' receptive vocabulary knowledge prior to or at the start of formal L2 English instruction (De Wilde et al., 2020a; Hannibal Jensen, 2017; Leona et al., 2021; Puimège & Peters, 2019; Sylvèn & Sunqvist, 2012). Studies investigating receptive vocabulary learning in this population have shown that words which are more frequent, more concrete and learned at a younger age by L1 speakers of English are easier to learn for young L2 English learners (De Wilde et al., 2020b; Puimège & Peters, 2019). Furthermore, these studies demonstrated that young L2 English learners also heavily rely on their L1, as cognates were known better than non-cognates. Only few studies also investigated learners' productive vocabulary knowledge (Bollansée et al., 2020; Sylvèn & Sundqvist, 2012). These studies tested form recall. Both studies showed that learners who engaged with English through out-of-school exposure were able to use more words productively compared to their peers who received less exposure through extramural English activities, even prior to the start of formal instruction. It remains to be seen however if and to which extent this learner population can use this vocabulary actively in a writing task.

2.2.Measuring lexical characteristics of L2 writing

It has been established that vocabulary is an important predictor of proficiency in all four skills (Milton, 2013). Many studies looking into L2 writing have investigated syntactic

measures (Bulté & Housen, 2014; Ortega, 2003) but researchers have also pointed out the importance of investigating lexical aspects in L2 writing in order to capture learners' proficiency in L2 writing (Skehan, 2009). An overview of studies investigating lexical aspects in writing is given below. I will also discuss the few studies that have investigated L2 English writing in young learners.

2.2.1. Complexity

Aspects which try to capture how lexically advanced a learners' productive language is, are categorized under lexical complexity. Lexical complexity research investigates both text-internal and text-external aspects of L2 writing. A text-internal aspect that has been widely investigated is lexical diversity, which measures the variety of words used in a text. A text-external aspect that has often been investigated is how frequent the words in the writing task are in a reference corpus. Text-external aspects are often referred to as aspects of lexical sophistication (Bulté & Housen, 2014; Kyle & Crossley, 2015; Kyle et al., 2018). Below, I will discuss more elaborately why it is important to look into diversity and lexical sophistication and how they can be measured.

Lexical diversity

As mentioned above, lexical diversity refers to the variety of words used in a text. Many studies have shown the importance of lexical diversity as a predictor of L2 proficiency in writing in different types of text such as adults' freewrites (Crossley et al., 2011), or argumentative texts and letters written by adolescent learners of French (Bulté and Housen, 2009). Bulté & Housen (2014), who investigated intermediate and advanced learners' L2 English writing, found mixed results. The measure they used for lexical diversity (D) did not significantly correlate with proficiency score. The Guiraud Index, often used as a measure of

lexical diversity, which in this study was considered to be a measure of lexical richness, did significantly correlate with proficiency score. The authors describe lexical richness as a combination of diversity and productivity (number of words used). As productivity has been shown to correlate well with proficiency level, the significant correlation with the Guiraud Index was to be expected. Other studies (e.g., Maamuujav et al., 2021) did not find a significant effect of L2 diversity. Studies which have investigated lexical diversity in young learners' writing will be discussed in more detail below (cf. 2.4).

There are many different indices of lexical diversity which are based on the type-token ratio in the text. Some are simple type-token ratios, others are more advanced transformations of the type-token ratio. One of the reasons why so many different measures are available is because researchers have tried to develop indices of lexical diversity which are not dependent on text length (McCarthy and Jarvis 2010). Zenker and Kyle (2021) investigated which indices were appropriate to use with short L2 English written texts (50 - 200 words). The authors found that both MATTR (moving average type token ratio) and MTLD (measure of textual lexical diversity) were stable measures when analyzing short texts.

Lexical sophistication

Another aspect of lexical complexity is lexical sophistication, which refers to how advanced the language is that is used in the text. Traditionally, researchers have focused on frequency measures to describe lexical sophistication because more advanced learners tend to use more low frequency words (e.g., Laufer & Nation, 1995).

Recently, researchers have pointed out that more variables need to be investigated in order to describe lexical sophistication. Kyle and Crossley (2016) looked into lexical sophistication in two types of writing tasks, independent writing and source-based writing. They investigated a set of variables that might contribute to lexical sophistication. The first group of

variables that they added, were psycholinguistic variables as studies have shown that these also influence lexical proficiency (Crossley & McNamara, 2012; Guo et al., 2013). Examples of psycholinguistic variables are concreteness, meaningfulness, familiarity and age of acquisition. Secondly, they added word range or contextual diversity, i.e. the number of contexts in which a word occurs (Adelman et al., 2006). Words with a wider range are considered less sophisticated. The authors also investigated the frequency of multi-word expressions as studies have shown that L2 texts which contain more frequent multi-word expressions typically receive higher scores (Kyle & Crossley, 2015). Another area under investigation was the type of semantic development. Studies found that writing which contained words with more meanings (high polysemy scores) and words with few superordinates (low hypernymy values) tend to get lower scores (Crossley et al., 2011, Guo et al., 2013). Finally, they also investigated the role of academic language on writing scores. In the present study, I will also look into multiple aspects of lexical sophistication in a context which has not been widely investigated, namely L2 narrative writing in young learners who are at the start of formal instruction.

2.2.2. Accuracy

Housen and Kuiken (2009) consider accuracy to be the most transparent of the CAF-constructs as it measures deviances from the norm. Accuracy has been operationalised through many different measures, which are often expressed in terms of errors. Polio and Shea (2014) make a distinction between five different types of measures: holistic measures (lexical and syntactic), error-free units, number of errors, number of specific error types and measures that take into account the severity of the error. Liao (2020) looked into the development of lexical and syntactic accuracy in L2 Chinese writing. The author distinguishes four types of L2 written accuracy measures: combined accuracy, lexical accuracy,

morphological accuracy and syntactic accuracy. Aspects of lexical accuracy can be ratio of correct lexis, ratio of lexical error-free clauses, number of lexical errors by word category and number of specific types of lexical errors (e.g., spelling). Below, I will discuss how this has been operationalized in studies looking into the writing skills of young L2 learners.

2.2.3. Fluency

Apart from complexity and accuracy, researchers have also investigated fluency in writing. Fluency is often operationalised by measuring the total amount of words produced in a task or by measuring the number of words written per minute (Johnson, 2017; Michel, 2017). The number of words produced in a task is sometimes also referred to as lexical productivity and this variable tends to correlate well with proficiency score (Bulté & Housen, 2014). Fluency has been conceptualized as one of the three basic components of vocabulary knowledge (next to size and depth) by Daller et al. (2007), but fluency has not often been investigated in studies focusing on vocabulary (Schmitt, 2019). One study investigating L2 English learning in young learners that took fluency into account was the study by Pfenninger (2021). In her study looking into writing development of primary school learners, fluency was operationalised as written text length in tokens. Pfenninger (2021) found text length was a predictor of proficiency score (more proficient learners wrote longer texts).

2.2.4. Studies investigating lexical characteristics in young learners' L2 English writing

Even though several studies have investigated lexical characteristics in order to explain language development (cf. supra), only few studies looked into how lexical characteristics predict L2 writing proficiency in young learners.

Verspoor et al. (2012) investigated young Dutch learners' L2 English writing. The participants were in the first and third year of secondary school (aged 12 to 14 years old). Even though the participants had received some formal English instruction in primary school, the first-year students did not have much experience with L2 English writing. Therefore, the participants wrote about simple, personalized topics (their school or their latest holiday). The writing samples were between 25 and 200 words long. The texts received a holistic score based on the CEFR-descriptors which resulted in 5 levels (A1.1 to B1.2). The texts were hand-coded and 64 variables were calculated. The variables measuring lexical complexity were Guiraud's index (to measure lexical diversity), number of chunks, word length and a customized lexical frequency profile based on the corpus under investigation. The last three variables measure lexical sophistication. The authors found that Guiraud's index and the use of chunks were strong discriminators between proficiency levels, whereas word length and frequency were less useful to predict proficiency level. The authors also investigated accuracy by looking into errors that were made. They looked into many different types of errors such as lexical errors (e.g., use of the L1) and spelling errors (e.g., phonetic spelling). They found that the total amount of errors was a discriminator between the different proficiency levels.

Maamuujav et al. (2021) investigated syntactic and lexical features in Spanish-speaking secondary school students' L2 English academic writing. Even though these learners were not at the start of formal English instruction, the study will be discussed here as it is one of the few studies investigating young learners' writing and has looked at a wider range of aspects of lexical sophistication. The authors used the computational tool Coh-Metrix to investigate several measures of lexical complexity: MTLD as a measure of lexical diversity and word length, word frequency, age of acquisition, familiarity, concreteness, imageability and meaningfulness as measures of lexical sophistication. Results show significant correlations between proficiency scores and age of acquisition, concreteness, imageability and

meaningfulness. About 23% of the variance in the proficiency scores could be explained by the lexical variables. The authors did not investigate the impact of fluency nor accuracy.

Kyle et al. (2021) investigated Dutch 12-to-14-year-olds' writing tasks. The authors looked into the role of syntactic complexity and syntactic sophistication in learners' longitudinal writing development. Syntactic sophistication was measured by a number of frequency indices (verb frequency, unfilled verb argument construction frames and verb argument construction frames with particular main verb lemmas) and measures looking into strength of association. The results for the frequency indices showed that participants started using less frequent main verbs over time. The results for verb argument constructions are less clear. The study showed that for low-proficiency learners frequency effects tend to be stronger in isolated words than in combinations of words as single words might be more easily remembered especially at low proficiency levels.

The current study will contribute to existing research looking into learners' writing by investigating both broad and fine-grained lexical characteristics in a large sample of narrative writing tasks written by young learners who are at the very start of formal L2 English instruction. The specific aims and research questions are laid out below.

3. Research Questions

The research questions of the study are:

RQ1: What are the lexical characteristics of young Flemish learners' L2 English writing in a picture narration task at the start of formal language learning?

RQ2: Which lexical characteristics discriminate between different proficiency levels?

The first research question aims to map what characteristics 11-12-year-olds' narrative writing exhibits. This question is important from a language acquisition angle as it will inform us about the language gains made through out-of-school exposure and will give information about writing skills in young learners, a context in which not many studies have been undertaken. Secondly, the results are important for language teaching pedagogy as this study can inform teachers about young learners' writing skills and possible differences between learners at the start of formal classroom instruction.

With the second research question I want to expand research on lexical characteristics of L2 writing by studying a context and a group of learners which have not been studied often.

4. Method

4.1.Context

Belgium, the country where this study took place, has three official languages: Dutch, French and German. The study took place in Flanders, the Dutch-speaking part of Belgium. The educational context concerning foreign language learning in Flanders is somewhat different from many other countries in Europe (Enever, 2011). The first foreign language to be taught is French, one of the official languages in Belgium. French classes typically start in the two final years of primary school when children are ten or eleven years old. Flemish legislation allows for playful language activities in other foreign languages (also English) on a voluntary basis but at the time of writing there were only few primary schools that offered English in their curriculum. Formal English classes typically start in the first year of secondary school or even in the second year of secondary school when learners are 12 to 13 years old. It can thus be assumed that most participants in the study have not received any formal instruction in English prior to this study.

4.2.Participants

The participants in this study were 3168 learners who attended the first year of secondary school in Flanders between 2019 and 2021. The learners came from 45 different schools and were taught by 52 different English teachers. The participants did the writing task in September or October 2019, 2020, or 2021 respectively. They had had a maximum of 15 hours of English instruction prior to the test.

The participants were a subgroup of the learners who did the writing activity in www.starttoetsengels.be (see 4.3. below). In total 25591 writing tasks were submitted in the tool, of which 9714 tasks were scored. I selected all the texts written by learners in the first year of secondary school which received a score. No personal information from the participants was gathered.

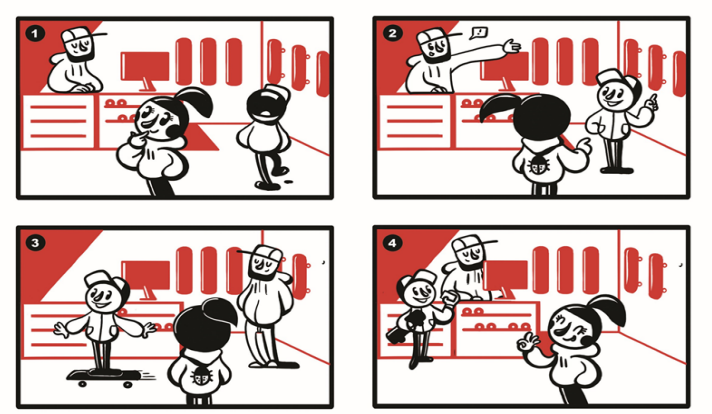
4.3.Instruments and procedure

4.3.1. Learner corpus

The participants all wrote a narrative writing task which was based on a picture story (see Figure 1). The writing task was part of a test developed to map Flemish leaners' prior knowledge of English at the start of formal English lessons. The test can be found online via www.starttoetsengels.be. This type of writing activity was added to the test as it was deemed suitable for young learners and for beginners as they can rely on the pictures to write their story. The test is free and easy to use. Teachers have to create a profile and create classes. For each class, they receive a link they can share with their pupils. They can test all four skills; listening and reading skills are corrected automatically, written and spoken texts are rated by the teachers.

Figure 1

Picture story used for the picture narration task (www.starttoetsengels.be, Artevelde University College Ghent)



In order to have a reliable scoring procedure, the rating in the tool www.starttoetsengels.be is done with benchmark texts which are supplemented by descriptors. The benchmark texts were developed through a two-stage approach following Humphry & Heldsinger (2020). Four texts were selected that were considered to be a good representation of each proficiency level (pre-A1, A1, A2, above A2). A fifth level was distinguished for learners who were not able to produce any English output (no output). As suggested by Humphry and Heldsinger (2020), the benchmark texts were supplemented with descriptors. The benchmark texts and descriptors can be found in the project *L2 English Writing in Young Learners* in the OSF repository (De Wilde, 2022). The tasks analysed in the current study were scored with the benchmark texts by 52 English teachers who used the test in their classroom. A quarter of the teachers who rated the writing tasks had followed a training given by the test developers on how to score the texts. The other teachers were either trained by a colleague who followed the initial training or used the guidelines on scoring the writing tasks which are available in the manual on the website (www.starttoetsengels.be). To check reliability of the teachers' scores, 319 writing tasks (every tenth task) were also scored

by an expert rater. The correlation between the teacher ratings and the expert ratings was high (r = .78).

In order to be able to calculate measures of lexical diversity and lexical sophistication in beginner-level written texts, all texts were checked for spelling errors manually by a researcher and afterwards doublechecked with an automatic spellchecker in Excel. Texts which were entirely written in Dutch or which only consisted of non-existing words, were assigned a no-output score, even if they had originally been given a pre-A1 score. The corrected texts were then converted to txt-files. This was done to be able to do automatic analyses for lexical diversity and lexical sophistication.

4.3.2. Measures of lexical diversity, lexical sophistication, accuracy and fluency

Lexical diversity was computed with the tool for the automatic analysis of lexical diversity (TAALED; Kyle et al., 2021). The tool calculates different measures of lexical diversity. For this study I calculated the Measure of Textual Lexical Diversity (MTLD; McCarthy & Jarvis, 2010). This was done because MTLD is stable across different text lengths (contrary to, for example, type-token ratio) and MTLD can be used for short texts (Zenker & Kyle, 2021).

Measures of lexical sophistication were calculated with the tool for the automatic analysis of lexical sophistication (TAALES; Kyle et al., 2018). With this tool it is possible to compute over 400 indices of lexical sophistication. An overview of the measures selected for this study can be found in Table 1. When various measures were available, I based the choice on findings from previous studies. Frequency measures to calculate mean frequency score (for all words, content words and function words in the texts) were taken from the Subtlex-US corpus (Brysbaert & New, 2009). This is a corpus based on subtitles. Previous studies have shown that such a corpus is appropriate for investigating language learning which has happened outside de classroom (De Wilde et al., 2020b; Puimège and Peters, 2019). Mean

range score was also calculated based on the Subtlex-US corpus. The psycholinguistic variables measured with this tool (concreteness, imageability, meaningfulness, familiarity) come from the MRC Database (Coltheart, 1981). Values from this database range from 100 to 700. The tool also offers an alternative for concreteness measures from the MRC Database, namely the concreteness ratings gathered by Brysbaert et al. (2014). As this database consists of more words and these ratings were also used in previous studies concerning word learning in this context and age group (De Wilde et al., 2020b; Puimège & Peters, 2019), I decided to use these ratings in my analyses. The concreteness values range from 1 to 5.

I used two measures that give an indication of the texts' accuracy. First, TAALES also computes index coverage for each measure. Since the texts were written by beginners, they sometimes contained English-like non-words, Dutch words and sometimes also French words. In order to be able to take this into account, I added the index coverage value for Subtlex-US, a measure which describes how many of the words in the learners' writing were present in the Subtlex-US corpus. Second, I counted the spelling errors made at the word level (phonetic spelling, using Dutch spelling rules, similar words such as to – two, and other spelling mistakes). This was done during the manual text correction. As a final measure, I added the word count calculated in TAALES. This measure is sometimes used as a fluency measure (e.g., Pfenninger, 2021). As the tasks in the present study were not timed, word count cannot be considered a measure of fluency but the measure still shows the extent to which learners were able to formulate their ideas in L2 English (lexical productivity).

Table 1

Lexical measures and their operationalization

Variable	Operationalization
Lexical diversity	Measure of Textual Lexical Diversity (MTLD; McCarthy & Jarvis, 2010)

Frequency	Subtlex-US corpus (Brysbaert & New, 2009), log-transformed frequency values
Range	Subtlex-US corpus (Brysbaert & New, 2009), log-transformed range values
Psycholinguistic variables: Imageability Meaningfulness Familiarity	MRC Database (Coltheart, 1981)
Psycholinguistic variables: concreteness	Brysbaert et al. (2014)
Age of acquisition ratings	Kuperman et al. (2012)
Multi-word frequencies (bigram, trigram)	COCA spoken corpus (Davies 2009), log-transformed frequency values
Polysemy	Wordnet database (Fellbaum, 1998)
Hypernymy	Wordnet database (Fellbaum, 1998)
Use of L2 English	TAALES index coverage value subtlex-US
Spelling errors	Manual text correction. Spelling errors at the word level (phonetic spelling, using Dutch spelling rules, similar words such as to – two, and other spelling mistakes)
Lexical productivity	TAALES word count

4.4.Analysis

I calculated descriptive statistics for each measure to be able to answer the first research question. In order to find an answer on research question two, Pearson correlations were calculated between lexical characteristics and the holistic scores on the writing tasks. I then ran a multiple regression model to get insights into the relationships between lexical measures and holistic proficiency scores. The model was built using best subsets regression. I first ran the full model with proficiency score as the dependent variable and 14 independent variables: MTLD, Subtlex-US frequency all words, meaningfulness, imageability, familiarity, concreteness, age-of-acquisition, bigram frequency, trigram frequency, polysemy,

hypernymy, Subtlex-US coverage, spelling errors and word count. Then, the best model was selected using the dredge function of the MuMIn package (version 1.43.17, Nakagawa & Schielzeth, 2013) in R (version 4.1.2, R Core Team, 2021) with AIC as the selection factor. In this model (cf. Table 5) nine variables were retained of which eight significantly contributed to the model. There was no multicollinearity between the independent variables: VIFs ranged between 1.07 and 3.07. The residuals follow a normal distribution. The relative importance of each variable in the model is expressed with the lmg-measure which was calculated using the calc.relimp function from the relaimpo package in R (version 2.2-6, Grömping, 2006). Raw data and analyses files are available in the OSF repository (De Wilde, 2022).

5. Results

5.1.Descriptive statistics

As mentioned in the method section, the texts were assigned a holistic score based on comparison with benchmark texts and descriptors. All 3168 narrative writing tasks received a score ranging from 'no output' to 'above A2'. The 'no output' score was given to 160 texts, which contained no English output. They were either written in Dutch or they consisted of (English-like) non-words. Most texts received a pre-A1 score (n=1170) or an A1 score (n=1200), 523 texts were scored with an A2-label and finally 115 texts were considered to be above A2-level. The shortest answer contained one word, the longest text was 182 words, the mean was 31 words. Two tasks were empty, they were left out of the analyses.

Descriptive statistics for all measures can be found in Table 2. In Table 3, the mean per proficiency level is given. A visual representation of the scores per proficiency level can be found in Appendix A (Figures 1 to 17).

Table 2

Descriptive statistics written texts

	Min	Max	Median	Mean	SD
1. MTLD	0	135.52	21.93	22.93	13.23
2. Subtlex-US frequency all words	0	5.68	4.84	4.70	0.59
3. Subtlex-US frequency content words	0	5.66	4.05	3.96	0.57
4. Subtlex-US frequency function words	0	6.33	5.82	5.66	0.99
5. Subtlex-US range all words	0	3.92	3.63	3.53	0.39
6. Meaningfulness	0	587	377	377.1	51.50
7. Imageability	0	634	348.9	350.7	51.14
8. Familiarity	0	645	607.3	600.4	60.56
9. Concreteness	0	4.85	2.74	2.76	0.48
10. Age of acquisition	0	14.67	4.46	4.55	0.77
11. Bigram frequency COCA spoken all words	-0.23	3.33	1.41	1.35	0.43
12. Trigram frequency COCA spoken all words	-0.26	2.44	0.45	0.45	0.37
13. Polysemy	0	30.00	8.94	9.02	2.86
14. Hypernymy	0	7.00	1.82	1.80	1.20
15. Coverage Subtlex-US	0	1.00	0.96	0.92	0.12
16. Spelling errors	0	37.00	3.00	4.03	4.27
17. Word count	1	182.00	27.00	30.55	17.56

The results in Table 3 show that the mean scores for certain variables visibly increased per level. This was clearly the case for word count, MTLD, and trigram frequency. The percentage of words in the writing task that was present in the Subtlex-US corpus also clearly increased. For the ‘no output’ category coverage was low (mean = 52 percent), which was to be expected as this category consists of tasks with hardly any English output. The fact that coverage was still 52% might have to do with the high number of cognates between Dutch and English. For the other levels, coverage was higher and also increased across levels (from 91 to 96%). The number of spelling errors overall decreased with higher proficiency levels, apart from the lowest level. Here the number of spelling errors was low as there was hardly any English output in the texts that could be corrected. Overall, the measures from the lowest proficiency group seem somewhat different from the other levels which is unsurprising as this is the level which represents tasks written by young learners who were not able to produce

any English output. The boxplots in appendix A further show that with increasing proficiency there seems to be less variation between individual learners for most measures, but not for word count.

Table 3

Mean scores for lexical characteristics per proficiency level

	No output	Pre-A1	A1	A2	Above A2
1. MTLD	4.81	20.15	24.77	28.10	33.28
2. Subtlex-US frequency all words	2.72	4.74	4.85	4.86	4.79
3. Subtlex-US frequency content words	2.42	3.98	4.09	4.05	4.05
4. Subtlex-US frequency function words	2.99	5.77	5.82	5.80	5.75
5. Subtlex-US range all words	2.20	3.55	3.63	3.64	3.63
6. Meaningfulness	326.4	384.5	378	374.6	372.1
7. Imageability	333.4	358.6	348.9	344.3	342.7
8. Familiarity	481.5	606.4	607.3	606.3	603.7
9. Concreteness	2.93	2.90	2.73	2.69	2.69
10. Age of acquisition	5.91	4.51	4.46	4.47	4.56
11. Bigram frequency COCA spoken all words	0.26	1.39	1.42	1.41	1.41
12. Trigram frequency COCA spoken all words	0.00	0.40	0.50	0.54	0.57
13. Polysemy	5.25	8.98	9.30	9.48	9.62
14. Hypernymy	4.79	4.98	4.77	4.65	4.71
15. Coverage Subtlex-US	0.52	0.91	0.95	0.96	0.96
16. Spelling errors	0.06	5.28	4.17	2.60	1.85
17. Word count	19.81	22.71	31.31	42.87	61.15

5.2.The relationship between lexical characteristics and proficiency score

Table 4 shows the correlations between the different lexical characteristics. The correlation between Subtlex-US frequency all words and Subtlex-US range is very high (r = .98). Therefore, it was decided to only add Subtlex-US frequency all words in the model. I opted for frequency rather than range as more studies have been done that looked into the role of frequency and this would make it possible to compare between the present study and previous findings. Furthermore, as the texts are relatively short, I chose one overall frequency measure

to be used in the regression model. The frequency measure containing all words also correlates more strongly with the writing score than separate frequency measures for content and function words (cf. Table 5). There were a few measures with a correlation higher than .70. I added those into the model but inspected in the final model whether there were no issues of multicollinearity by calculating the variance inflation factor (VIF), using the vif-function from the car-package in R (version 3.0 – 12, Fox et al., 2007).

In Table 5, the correlations between the proficiency score and the lexical measures are shown. The strongest relationship was found between score and word count (r = .52). Moderate correlations were found between proficiency score and coverage (r = .46), which is an accuracy measure and proficiency score and Subtlex-US frequency all words (r = .41), Subtlex-US frequency content words (r = .32), range (r = .43) and MTLD (r = .37). These four measures were used to investigate lexical complexity. All other measures showed a weaker but still significant relationship with proficiency score apart from meaningfulness which was not significantly correlated with proficiency score.

Table 4

Correlations between lexical measures for all written texts (n = 3166)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1.MTLD		.34***	.39***	.24***	.37***	.09***	-.02	.16***	-.08***	-.14***	.26***	.28***	.26***	-.18***	.29***	.10***	.35***
2.Subtlex-US frequency all words			.87***	.67***	.98***	.31***	.11***	.59***	-.12***	-.46***	.64***	.34***	.45***	-.16***	.72***	.20***	.21***
3.Subtlex-US frequency content words				.47***	.89***	.33***	.15***	.45***	-.05**	-.39***	.49***	.32***	.49***	-.27***	.56***	.24***	.23***
4.Subtlex-US frequency function words					.64***	.33***	.15***	.62***	-.06***	-.32***	.55***	.21***	.25***	-.03	.53***	.15***	.13***
5.Subtlex-US range all words						.37***	.18***	.61***	-.07***	-.42***	.57***	.31***	.45***	-.13***	.73***	.20***	.23***
6.Meaningfulness							.79***	.71***	.38***	-.09***	.09***	-.01	.21***	.05**	.22***	.08***	.01
7. Imageability								.63***	.53***	-.01	-.05**	-.12***	.07***	.28***	.13***	.04*	-.04*
8.Familiarity									.20***	-.17***	.36***	.13***	.27***	.10***	.46***	.10***	.11***
9.Concreteness										.23***	-.13***	-.10***	-.04*	.25***	-.04*	.01	-.07***
10.Age of acquisition											-.30***	-.13***	-.20***	.09***	-.35***	-.11***	-.04*
11.Bigram frequency												.47***	.30***	-.08***	.49***	.18***	.17***
12.Trigram frequency													.26***	-.19***	.25***	.11***	.23***
13.Polysemy														-.30***	.29***	.17***	.19***
14.Hypernymy															.00	-.05**	-.07***
15.Coverage Subtlex-US																.09***	.17***
16.Spelling errors																	.21***
17.Word count																	

* p < .05, ** p < .01, *** p < .001

Table 5

Correlations between lexical measures and overall proficiency score (n=3166)

	Score
1. MTLD	.37***
2. Subtlex-US frequency all words	.41***
3. Subtlex-US frequency content words	.32***
4. Subtlex-US frequency function words	.28***
5. Subtlex-US range all words	.43***
6. Meaningfulness	.03
7. Imageability	-.06***
8. Familiarity	.20***
9. Concreteness	-.13***
10. Age of acquisition	-.19***
11. Bigram frequency COCA spoken all words	.28***
12. Trigram frequency COCA spoken all words	.26***
13. Polysemy	.20***
14. Hypernymy	-.08***
15. Coverage Subtlex-US	.46***
16. Spelling errors	-.12***
17. Word count	.52***

* $p < .05$, ** $p < .01$, *** $p < .001$

I then built a multiple regression model using best subsets regression. The model (cf. Table 6) accounted for 50% of the variance in the proficiency scores. MTLD, word frequency, trigram frequency, Subtlex-US coverage and word count positively impacted the proficiency score. This means that more lexically diverse texts, texts containing more frequent words, texts containing more frequent trigrams, text which contained more words that were present in the Subtlex-US corpus and longer texts received a higher score. Imageability, age-of-acquisition and spelling errors negatively predicted the proficiency score. This means that texts containing words which are less imageable and acquired earlier by L1 speakers received a higher score, as did texts which contained fewer spelling mistakes. We also calculated lmg to

be able to comment on the relative contribution of each measure to the model. The total variance explained in the model was .50. Broad measures are relatively more important in the model than fine-grained measures. Twenty percent of the variance is explained by productivity (word count), 15% by accuracy (coverage and spelling), 5% is explained by lexical diversity and 5% by frequency. The other significant variables only explain a small amount of the variance in the model. Trigram frequency has a relative importance of .02 and is no longer significant after Bonferroni correction for multiple comparisons.

Table 6

Results of the regression model.

Predictors proficiency score	B	SE	β	sig	sig with correction	lmg
(Constant)	0.070	0.17	0.00	.681		
MTLD	0.007	0.00	0.11	.000***	***	.05
Subtlex-US Frequency all words	0.166	0.03	0.11	.000***	***	.05
Imageability	-0.001	0.00	-0.08	.000***	***	.01
Age of Acquisition	-0.049	0.02	-0.04	.003**	*	.01
Bigram Frequency	-0.057	0.04	-0.03	.130		.02
Trigram Frequency	0.096	0.04	0.04	.007**		.02
Subtlex-US coverage	2.200	0.13	0.30	.000***	***	.10
Spelling errors	-0.059	0.00	-0.28	.000***	***	.05
Word Count	0.024	0.00	0.45	.000***	***	.20
Model summary		Adjusted R-squared: .50, df 3156				

*** p<.001, **p<.01, *p< .05 / sig with correction: significance after Bonferroni correction for nine comparisons ***p<.0001, **p<.001, *p<.006

6. Discussion

The first research question inquired about the characteristics of young learners' L2 narrative writing at the start of formal instruction. The holistic scores show large differences between the learners, a similar result to what was found in other studies with young learners who had learned L2 English through out-of-school exposure (De Wilde et al., 2020a; Puimège & Peters, 2019). Most learners seemed to be able to write a story at the pre-A1 or A1 levels

(n=2370), showing that most of the learners had already picked up some English prior to the start of formal instruction and that they could use the language they had picked up in a narrative writing task. Only 5% of the learners (n = 160) were not able to produce any English in this task. 17% of the learners wrote a task which was given an A2-score and 4% received an 'above A2' score. This means that, based on this single task, 20% of the learners can write at the A2-level at the very start of formal instruction, confirming the earlier findings on learning English through out-of-school exposure. When looking at the lexical aspects of the texts, I also observed considerable differences between the learners. An inspection of the boxplots further showed less variation between more proficient learners. The language of more proficient learners seems to be more alike. This could point to the fact that they can communicate more efficiently.

I then investigated which lexical measures predicted the holistic score. The model explained 50% of the variance and showed that lexical aspects are important in L2 writing, at least at the early stages of language learning. Overall, the relative importance of broad measures was larger than that of more fine-grained measures but both types of measures significantly contributed to the model. The findings are in line with previous research (Milton, 2013). Similar to what Verspoor et al. (2012) found in their study with young learners, a number of broad measures predicted the writing score. These were total word count, MTLD, total number of spelling errors and percentage of English words in the text which were also present in the Subtlex-US corpus.

Word count, which shows learners' ability to formulate their ideas in English, was the strongest predictor of the proficiency score. This is unsurprising as these learners are at the very start of formal instruction and overall have a lower proficiency level and not all learners can easily express what they see in the picture prompt. Two other broad measures, the

percentage of English words and spelling errors also had a large impact on the proficiency score.

In the present study I also calculated measures of lexical complexity using NLP-tools. The two complexity measures that were most predictive of proficiency scores were MTLD (a measure of lexical diversity) and Subtlex-US frequency (a measure of lexical sophistication).

Lexical diversity was also a discriminator between proficiency levels in the study by Verspoor et al. (2012). The authors used a different measure of lexical diversity (Guiraud's index). In this study we used MTLD as this measure can be used with short texts (Zenker & Kyle, 2021) It must be noted however, that even though this is the most logical choice based on previous research, the reliability of the measure has not been investigated in texts which are shorter than 50 words. This warrants further investigation in future studies.

The findings further show that more proficient learners use more frequent words and words which were acquired earlier in life. These results contradict previous studies in L2 writing, which found that better writers used less frequent words (Laufer & Nation, 1995; Kyle & Crossley, 2015) and words which were acquired later in life (Guo et al., 2013). This could be due to the proficiency level of the learners in the present study. As the participants with the lowest proficiency scores were at the very start of English learning, they did not seem to be sensitive yet to variables such as word frequency and age of acquisition. The more 'advanced' learners (here at A2-level or slightly above) were able to use English in everyday contexts and the results thus show a frequency effect and an effect of age of acquisition. Writing tasks which received a higher score, reflected more typical use of everyday English. These findings are in line with findings on receptive vocabulary learning in learners with a similar profile (De Wilde et al., 2020b). Similar results were also found in a number of studies investigating spoken language (Berger et al., 2019; Crossley et al., 2019; Eguchi & Kyle, 2020). In these studies, it was shown that the learners used more frequent words over time.

Crossley and colleagues (2019) gave several explanations for these findings. They mentioned that function words, which are often high frequency words, might be difficult for beginners. The authors also hypothesized that early learners might not yet be attentive to frequency distributions of words. These hypotheses are in line with the findings in this study. The fact that we found this effect in writing might be explained by the type of writing task, a narrative task about an informal topic. The language necessary for this task is similar to the language used in the studies investigating naturalistic spoken discourse.

Some other fine-grained measures, tapping into aspects of lexical sophistication, also predicted holistic scores. These were trigram frequency and imageability. The results for imageability are in line with what was found in previous studies (e.g., Maamuujav et al., 2021) namely that more proficient writers use less imageable (and thus more sophisticated) words. More proficient writers also tended to use more frequent trigrams, a finding which is also in line with previous L2 writing research (Kyle & Crossley, 2015). It should be noted however that the contribution of these variables to the model was small.

The results of this study also have some consequences for the classroom. First, the study shows that, even though young learners mainly come across spoken input outside the classroom (De Wilde et al., 2020a; Peters, 2018; Puimège & Peters, 2019), some learners will be able to successfully use their acquired vocabulary knowledge to describe a picture story at the very start of L2 English lessons. At the same time there will be learners who do not know any English yet. When choosing suitable writing activities for their learners, teachers should thus take into account their learners' mixed abilities. This could for example be done by making the instruction more challenging for the more advanced writers or by providing key words for the beginners. Teachers could also ask more advanced writers to assist them in their writing classes and help improve the absolute beginners' writing skills e.g., through group work or peer feedback. Unfortunately, the variety in learners' proficiency and the need for

differentiation that stems from it is insufficiently addressed in commonly used textbooks in Flanders. Publishers could integrate these research findings in the materials they design to make it easier for teachers to effectively address the differences in proficiency in their classroom.

The results also clearly show that more proficient learners write longer texts. During writing classes teachers could practice writing fluency and productivity. As mentioned by Nation (2007) lower proficiency learners might profit a lot from having a number of useful sentences readily available. Teachers could integrate fluency activities in their lessons in order to practice writing longer texts. Another strong discriminator between proficiency levels is spelling. Teachers who teach English to beginners often stress that the first goal is to get the message across, and formal aspects are seen as less important. While it is important to stress that beginners' writing can and should not always be accurate, it would be good to also pay attention to spelling and provide activities in the classroom that train spelling or focus on strategies such as checking spelling through available sources or re-reading one's work. Finally, the study indicates that learners at this level are considered more proficient if they use more frequent words and trigrams. Teachers could teach everyday words and common expressions explicitly in the classroom in order to train their pupils' writing skills. When teachers differentiate in their writing classes, they can also address those aspects which are useful for learners at their current proficiency level e.g., a low proficient learner might want to train on productivity, while a learner with a higher proficiency might already be able to write longer texts without spelling mistakes and might profit more from expanding their vocabulary with multi-word combinations.

This study also has some limitations. First, as I wanted to investigate a large range of lexical characteristics, I was not able to take into account measures of syntactic complexity and sophistication and other aspects of accuracy. These aspects might explain additional

variation in proficiency scores. Second, the current study only concerned one type of writing and one single task. A different type of task might give different results with other lexical characteristics predicting proficiency (cf. Alexopoulou et al., 2017). Future studies could address these limitations.

7. Conclusion

The results of this study show large individual differences between young L2 English learners at the start of instruction. Lexical characteristics explain 50% of the variance between learners' writing scores. Some of these are broad measures such as word count, number of spelling errors, number of words present in an English corpus and a measure of lexical diversity. Additionally, there are also more fine-grained measures (frequency, imageability, age of acquisition) which predict the learners' writing scores. The study confirms the importance of vocabulary knowledge for writing in the early stages of L2 English learning.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

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: Task Effects in a Large-Scale Learner Corpus. *Language Learning*, *67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Berger, C. M., Crossley, S. A., & Kyle, K. (2019). Using Native-Speaker Psycholinguistic Norms to Predict Lexical Proficiency and Development in Second-Language Production. *Applied Linguistics*, *40*(1), 22–42. https://doi.org/10.1093/applin/amx005

Bollansée, L., Puimège, E., & Peters, E. (2020). “Watch out! Behind you is the enemy!” An exploratory study into the relationship between extramural English and productive vocabulary knowledge. In *Pop Culture in Language Education* (pp. 199-214). Routledge.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

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

Bulté, B., & Housen, A. (2009). The development of lexical proficiency in L2 speaking and writing tasks by Dutch-speaking learners of French in Brussels. In *Task-based language teaching conference*, Lancaster, England.

Bybee, J. L., & Hopper, P. J. (Eds.). (2001). *Frequency and the Emergence of Linguistic Structure* (Vol. 45). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.45

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., Ballon, P., & De Maeyer, S. (2018). An information system design theory for the comparative judgement of competences. *European Journal of Information Systems*, *27*(2), 248–261. https://doi.org/10.1080/0960085X.2018.1445461

Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication: Predicting L2 writing proficiency. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute Frequency Effects in Second Language Lexical Acquisition. *Studies in Second Language Acquisition*, *41*(04), 721–744. https://doi.org/10.1017/S0272263118000268

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and Assessing Vocabulary Knowledge*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667268

Davies, M. (2009). The 385+ million word *Corpus of Contemporary American English* (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

De Wilde, V. (2022, November 25). L2 English writing in young learners. https://doi.org/10.17605/OSF.IO/WDETY

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020a). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition*, *23*(1), 171–185. https://doi.org/10.1017/S1366728918001062

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020b). Learning English Through Out-of-School Exposure: How Do Word-Related Variables and Proficiency Influence Receptive Vocabulary Learning? *Language Learning*. https://doi.org/10.1111/lang.12380

Eguchi, M., & Kyle, K. (2020). Continuing to Explore the Multidimensional Nature of Lexical Sophistication: The Case of Oral Proficiency Interviews. *The Modern Language Journal*, *104*(2), 381–400. https://doi.org/10.1111/modl.12637

Ellis, N. C. (2002). Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, *24*(02). https://doi.org/10.1017/S0272263102002024

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: an eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. https://doi.org/10.1017/S0272263117000109

Enever, J. & British Council. (2011). *ELLiE - Early language learning in Europe*. British Council.

Fellbaum, C. (1998). A Semantic Network of English: The Mother of All WordNets. In P. Vossen (Ed.), *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 137–148). Springer Netherlands. https://doi.org/10.1007/978-94-017-1491-4_6

Fox, J., Friendly, G. G., Graves, S., Heiberger, R., Monette, G., Nilsson, H., ... & Suggests, M. A. S. S. (2007). The car package. *R Foundation for Statistical Computing*.

Goodier, T., & Szabo, T. (2018) *Collated representative samples of descriptors of language competences developed for young learners.* Eurocentres

Grömping, U. (2006). Relative Importance for Linear Regression in *R*: The Package relaimpo. *Journal of Statistical Software*, *17*(1). https://doi.org/10.18637/jss.v017.i01

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

Hannibal Jensen, S. (2017). Gaming as an English Language Learning Resource among Young Children in Denmark. *CALICO Journal*, *34*(1), 1–19. https://doi.org/10.1558/cj.29519

Housen, A., & Kuiken, F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Hulstijn, J. H. (2003). Incidental and Intentional Learning. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 349–381). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756492.ch12

Humphry, S., & Heldsinger, S. (2020). A Two-Stage Method for Obtaining Reliable Teacher Assessments of Writing. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.00006

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, *37*, 13–38. https://doi.org/10.1016/j.jslw.2017.06.001

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring Longitudinal Writing Development Using Indices of Syntactic Complexity and Sophisticiation. *Studies in Second Language Acquisition*, *43*(4), 781–812. https://doi.org/10.1017/S0272263120000546

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Lefever, S. (2010). *English skills of young learners in Iceland: "I started talking English when I was 4 years old. It just bang… just fall into me."*. Menntakvika Conference, Reijkjavik, Iceland.

Leona, N. L., van Koert, M. J. H., van der Molen, M. W., Rispens, J. E., Tijms, J., & Snellings, P. (2021). Explaining individual differences in young English language learners' vocabulary knowledge: The role of Extramural English Exposure and motivation. *System*, *96*, 102402. https://doi.org/10.1016/j.system.2020.102402

Liao, J. (2020). Do L2 lexical and syntactic accuracy develop in parallel? Accuracy development in L2 Chinese writing. *System*, *94*, 102325. https://doi.org/10.1016/j.system.2020.102325

Maamuujav, U., Olson, C. B., & Chung, H. (2021). Syntactic and lexical features of adolescent L2 students' academic writing. *Journal of Second Language Writing*, *53*, 100822. https://doi.org/10.1016/j.jslw.2021.100822

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In *The Routledge handbook of instructed second language acquisition* (pp. 50-68). Routledge.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist & B. Laufer (eds.) *L2 vocabulary acquisition, knowledge*

and Use: New Perspectives on Assessment and Corpus Analysis. 57-78. Eurosla Monograph Series 2.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nation, P. (2007). The Four Strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt039.0

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

Peters, E. (2018). The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, *169*(1), 142–168. https://doi.org/10.1075/itl.00010.pet

Pfenninger, S. E. (2021). About the INTER and the INTRA in age-related research: Evidence from a longitudinal CLIL study with dense time serial measurements. *Linguistics Vanguard*, *7*(s2), 20200028. https://doi.org/10.1515/lingvan-2020-0028

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27. https://doi.org/10.1016/j.jslw.2014.09.003

Puimège, E., & Peters, E. (2019). Learners' English Vocabulary Knowledge Prior to Formal Instruction: The Role of Learner-Related and Word-Related Variables. *Language Learning*, *69*(4), 943–977. https://doi.org/10.1111/lang.12364

R Core Team (2021). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(02), 261–274. https://doi.org/10.1017/S0261444819000053

Skehan, P. (2009). Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary*. (Unpublished doctoral dissertation). Karlstad University Studies, Karlstad, Sweden.

Sylvén, L. K., & Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL*, *24*(03), 302–321. https://doi.org/10.1017/S095834401200016X

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

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Zhang, R., Zou, D., Cheng, G., Xie, H., Wang, F. L., & Au, O. T. S. (2021). Target languages, types of activities, engagement, and effectiveness of extramural language learning. *PLOS ONE*, *16*(6), e0253431. https://doi.org/10.1371/journal.pone.0253431

Appendix A: Boxplots showing the scores for the variables measuring lexical complexity accuracy and fluency per proficiency level.

Figure 1

Boxplot showing the MTLD-scores per proficiency level.

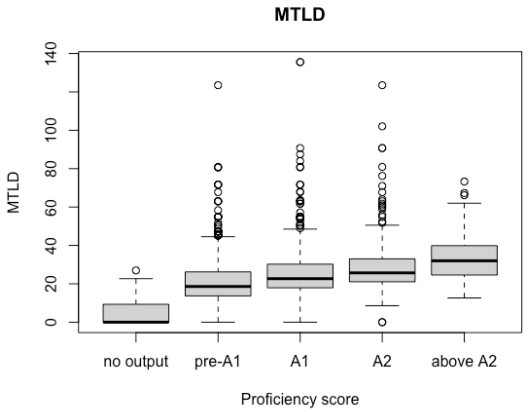


Figure 2

Boxplot showing the Subtlex-US frequency score (all words) per proficiency level.

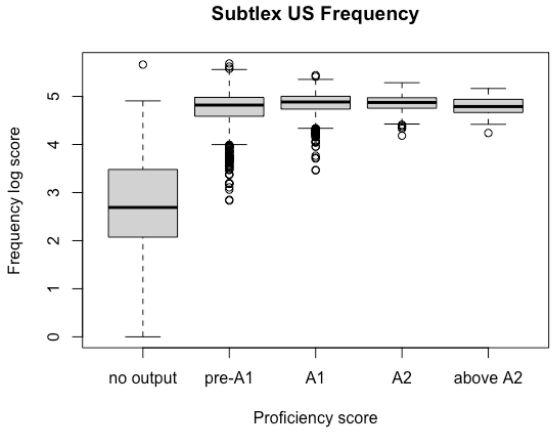


Figure 3

Boxplot showing the Subtlex-US frequency score (content words) per proficiency level.

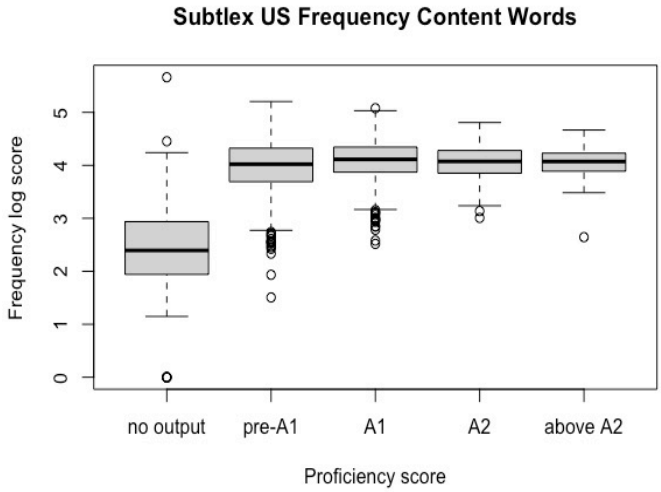


Figure 4

Boxplot showing the Subtlex-US frequency score (function words) per proficiency level.

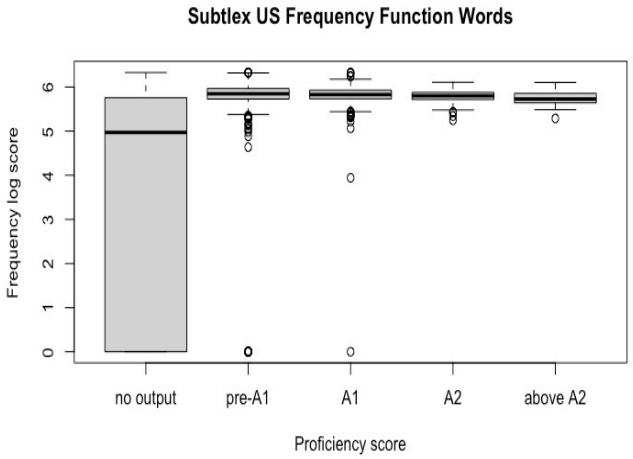


Figure 5

Boxplot showing the Subtlex-US range score (all words) per proficiency level.

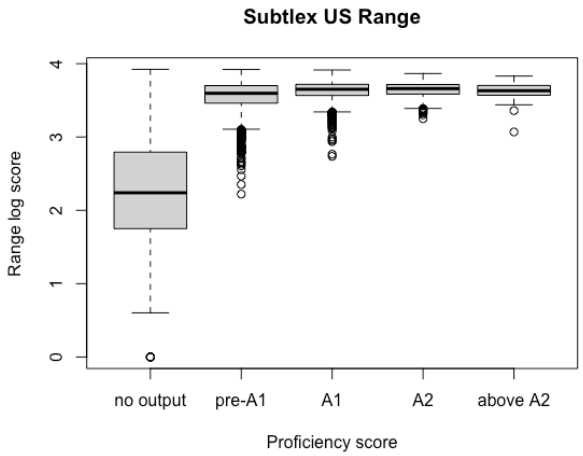


Figure 6

Boxplot showing the Meaningfulness score per proficiency level.

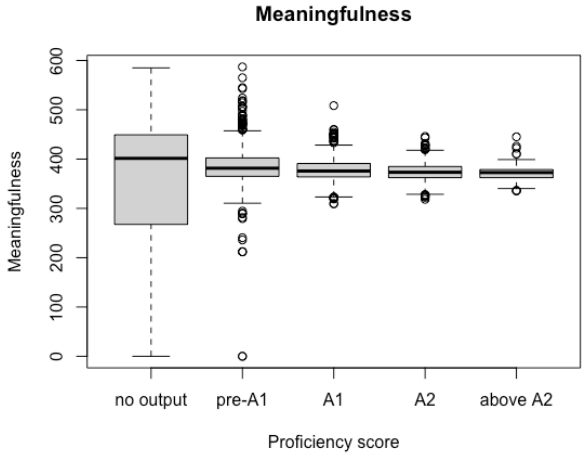


Figure 7

Boxplot showing the Imageability score per proficiency level.

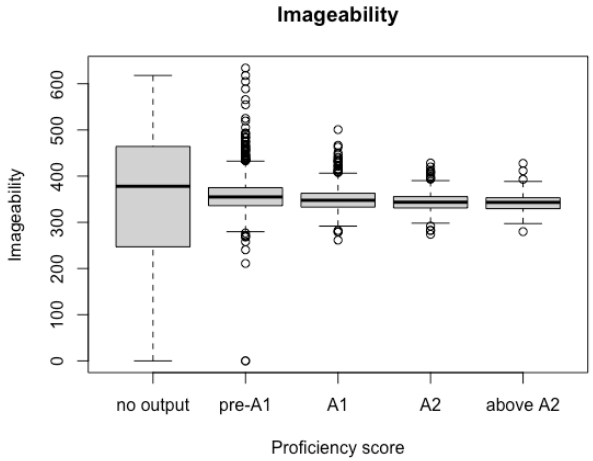


Figure 8

Boxplot showing the Familiarity score per proficiency level.

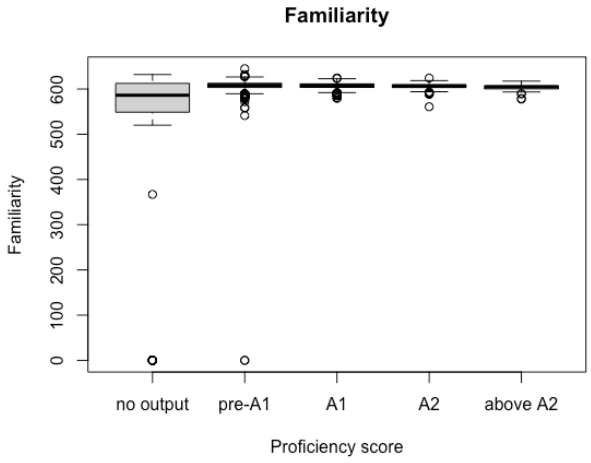


Figure 9

Boxplot showing the concreteness score per proficiency level.

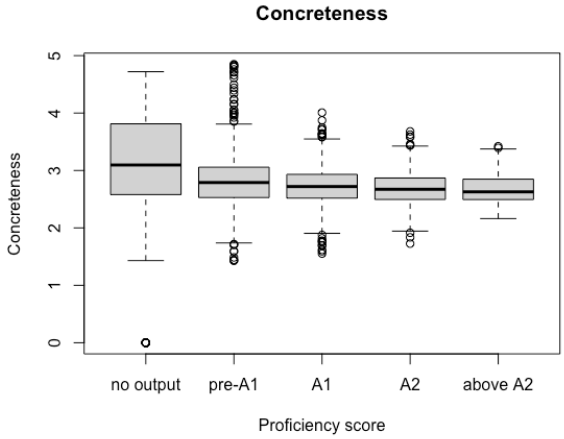


Figure 10

Boxplot showing the Age of Acquisition score per proficiency level.

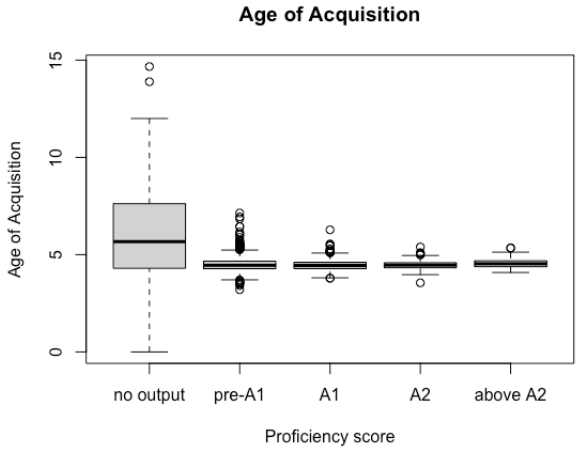


Figure 11

Boxplot showing the Bigram frequency score per proficiency level.

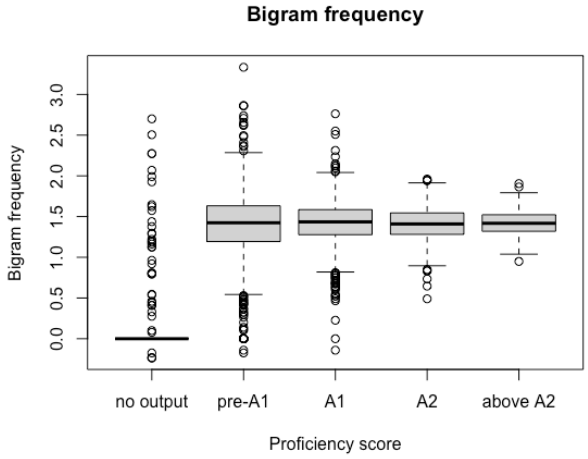


Figure 12

Boxplot showing the Trigram frequency score per proficiency level.

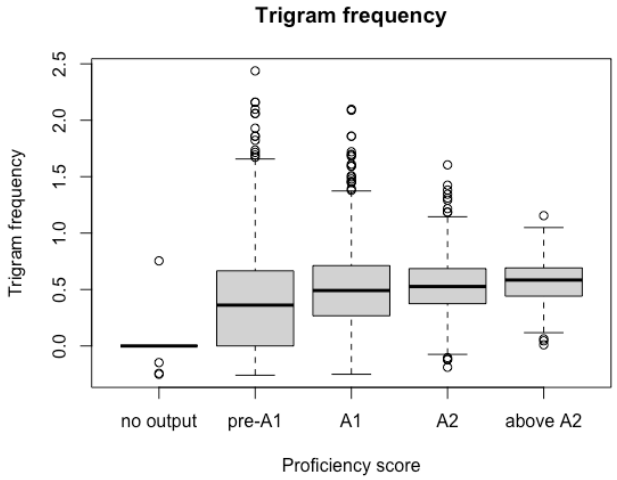


Figure 13

Boxplot showing the Polysemy score per proficiency level.

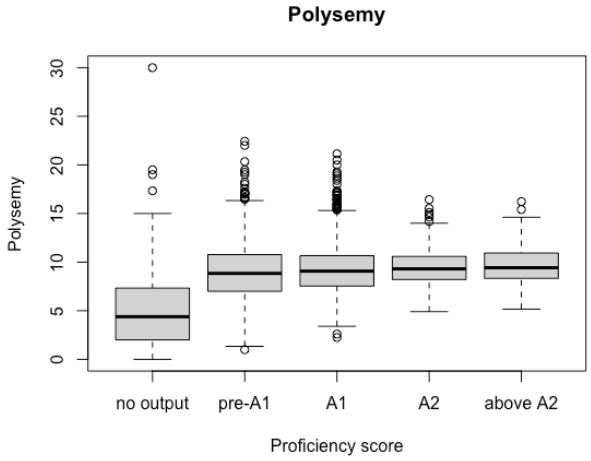


Figure 14

Boxplot showing the Hypernymy score per proficiency level.

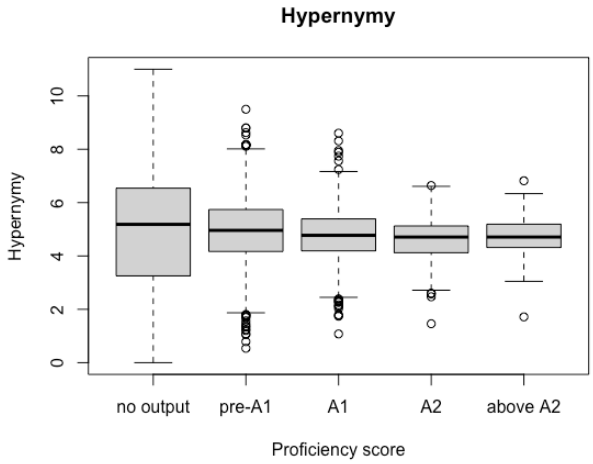


Figure 15

Boxplot showing the Subtlex-US coverage per proficiency level.

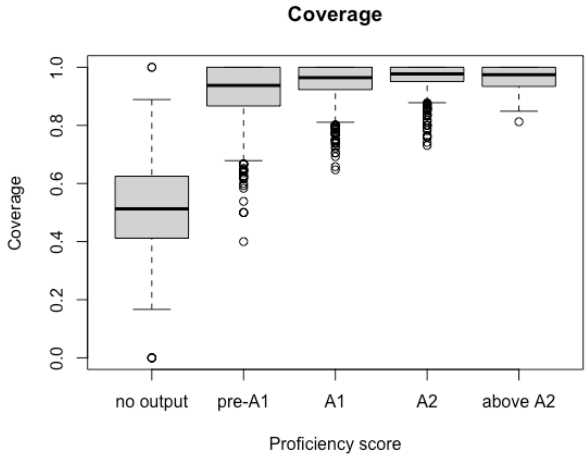


Figure 16

Boxplot showing Spelling errors per proficiency level.

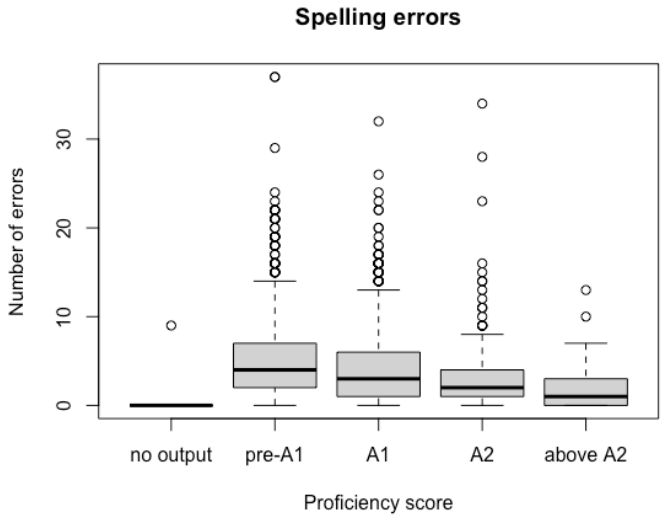


Figure 17

Boxplot showing Word count per proficiency level.