Evaluating Written L2-English Learner Texts Through Manual And Automated Quantitative Analyses

ANDREA ÁGNES REMÉNYI

Preprint, to appear in:
Dósa, A. & Magnuczné Godó, Á. &
Nagano, R. L. & Schäffer, A. (eds.)
Crossing boundaries: Space, identity and discourse in Anglophone studies.

Introduction

A recurring language testing and assessment question is how to conceptualise language proficiency levels according to the Common European Framework of Reference for Languages (CEFR, 2001) in quantifiable features of English grammar and vocabulary. In other words, what are the characteristics of a certain CEFR level, in terms of not the language skills but the patterns of its syntactic and lexical control and range (i.e., accuracy and complexity of both)? A limited, though not necessarily minimal, set of those quantifiable characteristics could inform the teaching-learning process and also facilitate spoken and written learner text evaluation.

As part of the validation process of the B2+ CEFR-level language examination (henceforth: exam) for English majors taken at the end of their first year of studies at a Hungarian university, we set out to detect the systematic patterns of syntactic and lexical characteristics of a written corpus and their match to B2+ expectations. The project research question is whether that exam measures English language proficiency at the B2+ level in a valid and reliable way, based on the patterns of syntactic and lexical complexity, as far as the written texts are concerned. The present research question is what quantitative method of analysis can help us select the most relevant features of complexity.

Thus, in this paper, two related issues are discussed. On the one hand, I will focus on the problem of the possibility of proving that a certain language exam measures what it intends to measure—this is called exam validation. On the other hand, I will approach the issue of the possibility of finding ways to automatically assess English-as-a-foreign-language (henceforth, and more generally: L2-English) written texts produced at language exams. More precisely, the following

research question will be scrutinised here: What quantitative method of analysis can help us select the most relevant distributive patterns of syntactic and lexical complexity features to characterise L2-English learner texts? At a later step of the research project, the best predictor variables to separate texts at or above the B2+ proficiency level from those below will be identified.

First, let me focus on the problem in the conceptualization of CEFR proficiency levels: are they quantifiable as far as grammar and vocabulary are concerned, for L2-English? As an example, an L2-English language exam will be introduced that needs validation. Two texts written at that exam will get a close look to indicate the efforts, both automated and manual, to find quantifiable features. Then the results produced through some multivariate analytical systems will be discussed and statistically compared. Such a systematic comparison across those systems will be recommended because some of the variables from each of them thus corroborate one another. Finally, I will suggest how the results can provide the basis for partial automated evaluation of those texts at language exams.

Background

Range and Control in the CEFR

The conceptualization of foreign language proficiency in terms of grammar and vocabulary has been the focus of research for some time. Efforts preceding the CEFR started in the 1970s with Van Ek (1975), who, in a project funded by the Council of Europe, first described a foreign language (FL) level of English called *The Threshold Level,* in terms of grammatical structures, for the sake of international applicability in course design, coursebook design and language testing across Europe. Directly preceding the work on the CEFR, that book was revised and re-published (Van Ek and Trim 1991a), followed by similar volumes on the Waystage and Vantage levels (Van Ek and Trim 1991b; Van Ek and Trim 2001). These three levels were later to be labelled B1 (Threshold), A2 (Waystage) and B2 (Vantage) in the CEFR.

The CEFR (2001) and the CEFR Companion Volume (2018) provide information on both the general linguistic range and vocabulary range in its can-do descriptors, on the one hand, and grammatical and vocabulary control, on the other. Because their framework of reference is not L2-English but any FL, language-specific information is absent from the can-do descriptors. While the B2+ level gets little attention in the CEFR, the neighbouring levels of C1 and B2 are useful starting points to conceptualise the general features of the level in between them. For example, the general linguistic range of the C1 and B2 levels is described by these can-do descriptors:

C1:
Can select an appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say.
B2:
Can express him/herself clearly and without much sign of having to restrict what he/she wants to say. […] Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so. (CEFR 2001, 110)

Thus, "a broad range" or "sufficient range" of language are emphasised, making it clear that both syntactic and vocabulary features are covered by these descriptors. Next, vocabulary range is described this way:

C1:
Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.
B2:
Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. (CEFR 2001, 112)

Here, "a broad lexical repertoire" or "a good range of vocabulary" are in the focus of the descriptions. As far as control is concerned, this is how, first, vocabulary control is described, stressing "minor slips" only in the case of the C1 level, while in the case of B2, some points of incorrect word usage do occur:

C1:
Occasional minor slips, but no significant vocabulary errors.
B2:
Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication. (CEFR 2001, 112)

And finally, grammatical accuracy is described with the help of the following can-do descriptors, with "rare" errors at C1 while "occasional", "non-systematic" errors at B2:

C1:
Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.
B2:
Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect. […] Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding. (CEFR 2001, 114)

The CEFR levels and their can-do descriptors are, obviously, immensely useful in the conceptualisation of the various foreign language proficiency levels. However, to find actual L2-English syntactic and lexical features that characterise learner texts at those proficiency levels as opposed to those below them, the guidance offered by the CEFR, due to its nature, is limited. To get a step ahead, let us first study further the concepts of "range" and "control".

Conceptualisations of Syntactic (and Lexical) Complexity

In recent theorisation and corpus-based analysis, "range" and "control" are often called "complexity" and "accuracy", respectively. In Bulté and Housen's model (2012, 2014), L2 proficiency is described by the triad of complexity, accuracy and fluency. Among them, complexity is defined as "the elaborateness, richness and diversity" of the learner's performance (Housen and Kuiken, 2009, 4). Syntactic complexity is a sub-component of that construct, belonging to linguistic complexity, and including sentence, clausal and phrasal levels of complexity. Lexical complexity is another sub-component of linguistic complexity, and includes collocational and lexemic complexity. In this section, for lack of space, only syntactic complexity is discussed in some detail.

Ortega (2003) considers syntactic complexity as a sign of syntactic maturity and defines it as "the range of forms that surface in language production and the degree of sophistication of such forms" (2003, 492). She emphasises that it is related to the learner's syntactic repertoire, represented in various features, including the "length of production units, amount of embedding, range of structural types and sophistication of the particular structures" (2003, 492).

Lu (2010, 2017) conceptualises syntactic complexity as a multidimensional construct, which can be viewed from two perspectives, both requiring distinct measures. The perspective of L2 testing and assessment considers its quality, while L2 writing research examines it from the variability perspective.

The above and other approaches seeking to quantify syntactic complexity in L2-English writing seem to belong to four research threads. In one of those threads, native speaker (NS) written texts are compared systematically to non-native (NNS) texts, with the latter often sub-grouped according to their writers' proficiency levels. For example, Ai and Lu (2013) examined the length of production units (the mean length of sentences/clauses/T-units), the amount of subordination (dependent clause per clause/per T-unit), the amount of coordination (coordinate phrases per clause/per T-unit, T-units per sentence), and the degree of phrasal sophistication (complex nominal per clause/per T-unit). Mancilla et al. (2015) used similar categories in their analysis.

In another thread of research, L2 development is examined longitudinally by comparing learners' written texts as their English proficiency has been developing. Some researchers in this thread employ similar measures of the length of various

syntactic units, of subordination and coordination, and of phrasal sophistication, including Polat et al. (2019) and Lei et al. (2023), and find that several of those measures are good indicators of syntactic complexity development.

In a third thread of research, the distinctive patterns and distribution of syntactic complexity are examined at various levels of L2-English proficiency (as assessed by human raters). For example, Taguchi et al. (2013) analysed texts on the basis of phrasal level measures (determiners, attributive adjectives, post-modifying prepositional phrases, etc.) and clausal level measures (subordinating conjunctions, *that*-relative clauses, etc.). Crossley and McNamara (2014), in their longitudinal research design, also concentrated on both the phrasal and clausal levels, and quantified, for example, modifiers per noun phrase, prepositional phrases and infinitives. All these features have been found by the respective researchers to be indicative of L2 proficiency.

The fourth research thread seeks to compare the register-, genre- and task-specific features of syntactic complexity in learner writing. For example, based on the long-established research tradition of register variation by Douglas Biber and his colleagues, Biber et al. (2016) analysed the distribution and co-occurrence of 23 grammatical features in texts produced in written, spoken and integrated tasks. They found that, for example, longer words, prepositional phrases, attributive adjectives, passive verb constructions and verb+*that*-clauses were significantly more frequent in written texts, and those tended to co-occur more frequently in texts receiving higher scores by human raters.

The present L2-English corpus-based research project aims to contribute to the findings of the above research threads. As explained below, our corpus data are analysed with the help of the various measures of syntactic and lexical complexity to identify the best predictive patterns of L2-English proficiency as far as lower level vs. higher level texts are concerned (around the B2+ level, in our case), on the one hand. And on the other, with the help of a statistical meta-analysis, the most robust measures are sought by comparing the various multivariate analytical systems.

Research Methods

Data Collection: The Language Exam and the Corpus

The research project is based on a growing corpus involving L2-English learner texts written at the Basic Language Exam (BLE). The BLE is a B2+ level proficiency-type assessment event at a Hungarian university, a high stakes exam for English majors at the end of their first year of studies, including students in both the BA (full-time, part-time) and the English teacher training programmes. The

stakes are high because the exam can be taken only twice in one's studies altogether: in case of a fail, only one re-take is possible in a following semester.

It is called "Basic" to indicate that it is the basis for those students' further studies. It is taken by around 150 candidates per year, in the autumn and spring semesters. Its written part, taken on one day, consists of a "Use of English" test followed by a reading and a writing component—the latter is in the focus of this project. The oral component takes place on another day. Presently, listening comprehension is not evaluated due to technical difficulties. Because of the fail rate, the exam is recurrently criticised for its severity. This is why exam validation is crucial, to prove that the level against which the exam assesses candidates' proficiency is indeed B2+ and not higher.

Table 1. Grammar and vocabulary descriptors for BLE written text production (PPCU 2017)

Grammar		Vocabulary	
5 pts.	wide range of structures, few inaccuracies that do not hinder/ disrupt communication	**5**	wide range of vocabulary, accurate vocabulary communicating clear ideas, relevant content
4	good range of structures, occasional inaccuracies hinder/disrupt communication	**4**	good range of vocabulary, occasionally inaccurate vocabulary communicating mainly clear ideas; overall relevant to content
3	limited range of structures, frequent inaccuracies hinder/disrupt communication	**3**	limited range of vocabulary, frequently inaccurate vocabulary communicating some clear ideas; occasionally relevant to content with some chunks lifted from prompt
2	*in between*	**2**	*in between*
1	no range of structures, mostly inaccurate	**1**	no range of vocabulary, mostly inaccurate vocabulary communicating few clear ideas; mostly irrelevant to content with several chunks lifted from prompt

The written text production component of the BLE takes 45 minutes to write and is based on two or three prompts in one of three genres in each exam period: formal letters, narratives and reviews. For example, a formal letter prompt was this:

Topic 1. You recently stayed in a motel in New Orleans. The weather was unusually hot for the time of the year and the air conditioning unit in your room did not work properly. Write a letter to the hotel manager. In your letter:
– give details of what went wrong
– explain what you had to do to overcome the problem while you stayed there
– say what action you would like the manager to take.
Write a 180-200 text body. (PPCU 2022)

Each text written by the candidates is evaluated by two blind reviewers against a four-category set of rating scales, including descriptors on task achievement, coherence and cohesion, grammar and vocabulary. The descriptors of the latter two are listed in Table 1 —as you can see there, the concepts of complexity ("range") and accuracy are present in describing both categories.

Presently, the BLE corpus contains 395 texts of approximately 200 running words each, and includes formal letters and narratives. All texts are collected following the BAAL recommendations on ethical research (BAAL 2021) and are used for research and publication purposes on the basis of the writers' written informed consent. The texts are hand-written at most BLE sittings. Those texts are carefully type-transcribed in two versions in a file saved with a code name only: a verbatim version keeping the idiosyncracies of the original, excluding the author's name (for the manual analysis), and another version where the spelling and punctuation are changed to follow the conventions of standard English (for the automated analysis).

In this paper the output from two exams will be analysed: in one of them, inquiry letter topics were offered (three topics to choose one from, May 2017, N=73), while in the other a complaint letter was to be written (two topics to choose one from, April 2022, N=98). This subcorpus contains altogether 35,145 running words.

Data Analysis

Human raters across the world evaluate hundreds of thousands of written candidate texts at L2-English language exams every year. Their evaluation work is impressionistic, due to time constraints, and supported from three sources to secure reliability: the raters' training and experience, the descriptors of the rating scales in that exam, and the cooperation among raters. Although syntactic and lexical complexity are only two of the components to be evaluated in their work, still, our research attempts to quantify features of those two complexity categories can also be pictured as uncovering and modelling that latent analytical effort by the human raters.

In our analysis of the BLE corpus, we have been employing four external multivariate automated analytical systems of syntactic and lexical complexity, coupled with our own automated and manual variables. The external analytical systems are the Multidimensional Analysis Tagger (MAT; Nini 2019, 2021), the

Web-based L2 Syntactical Complexity Analyzer (L2SCA; Lu 2017, Ai 2022), the CEFR-based Vocabulary Level Analyzer (CVLA 2023; Uchida and Negishi 2018) and Lextutor (Cobb 2023); most of them use the Stanford parser and the General Service List as the bases of their analysis. Our own variables are being developed, and presently include an automated (rest/K1 tokens) and several manual variables (verb tense/aspect distribution, article usage vs. countability, human raters' scores and accuracy measures)—the comparable quantification of those are still being developed. See Table 2 for an overview.

Table 2. **The multivariate analytical system employed in the BLE research project**

Syntactic complexity	Lexical complexity
Biber-tagger/MAT (Nini 2019, 2021): - Past Tense - Perfect Aspect - nominalisations - existential *there* - *be* as a main verb - agentless passives/*by* passives - various relative clause variables - necessity modals, etc.	**Biber-tagger/MAT:** - conjuncts (*moreover*) - downtoners (*almost, nearly*) - private verbs (*consider, realise*) - public verbs (*announce, explain*) - place adverbials, time adverbials - attributive, predicative adjectives, etc. (altogether 67 variables)
L2SCA (Lu 2017; Ai 2022): - mean length of clause/sentence/T-unit - amount of subordination (e.g. clause/sent.) - amount of coordination - degree of phrasal subordination, etc. (cca. 20 variables)	**Lextutor-based variables (Cobb 2023):** - family/type/token distributions - K1-K2-K3up distributions - type-token ratio - lexical density (content words/ tokens) (cca. 20 variables)
CVLA (2023; Uchida-Negishi 2018): - verbs/sentence - ARI, a readability measure	**CVLA:** - average lexical difficulty - BperA, the ratio of two frequency categories (altogether 4 variables)
Our own variables - manual: accuracy measures; verb tense/aspect distribution, article usage vs. countability; raters' points, etc.	**Our own variables - automated**: rest/K1 tokens (based on Lextutor)

The results yielded by the above multivariate analytical systems are standardised for comparability, and then further examined by inferential statistical testing. Presently, I am using correlation and factor analysis to reveal the internal relationships and the latent variables in the data matrix. Those calculations are also

useful to detect the correspondences among the variables across the employed multivariate analytical systems.

Results and Discussion

Automated Analysis: An Example of the CVLA Results

Due to space constraints, let me give here only two examples, two short glimpses, into the results of the analyses, based on the formal letters collected at two BLEs. The first one shows the automated analyser CVLA, a Japanese development to classify English texts into 12 CEFR levels: pre-A1, A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, B2.1, B2.2, C1, C2 (Uchida and Negishi 2018; CVLA 2023). It is based on four variables altogether: one measure of syntactic complexity, two measures of vocabulary frequency and a readability measure. The syntactic complexity variable calculates the number of verbs per sentence. One of the vocabulary frequency variables is average vocabulary difficulty, which assigns a CEFR-related value to each content word, based on the CEFR-J wordlist (developed hand-in-hand with the CVLA). The other vocabulary variable is based on the same wordlist, and calculates the ratio of B-level vs. A-level content words. The fourth variable is the ARI, a readability measure, which simply calculates the average ratios of character per word and word per sentence—as such, it is linguistically uninformed and may raise questions about its applicability. The scores based on the four variables are then converted into a CEFR category, plus an overall CEFR-category is also calculated by the programme.

In spite of the low number of the variables, the CVLA results look intuitively reliable. Also, the output results closely correspond to human rater classifications into CEFR levels. To illustrate, let me show how it categorises a high-rated (Text A) vs. a low rated text (Text B) written for the same BLE from the same prompt (see Topic 1 above); note that the verbatim version is shown here, while a corrected version was used as input for the automated analysis.

Text A

Dear Sir or Madam,

I am writing to you in regards to my recent stay at your motel located in New Orleans. Unfortunately I had a terrible experience during my stay. I would like to be compensated in one way or another, as I feel I have been misadvertised to.

This unpleasurable experience of mine came about because of a faulty air conditioning unit. As you know the weather during my stay was unusually hot, and the lack of air conditioning made my stay unbearable. To relieve some of the heat stress, I had to pay out of my own pocket to purchase a portable ventillator, which of course had to leave at your motel, as it was too big to take with myself when I had to leave.

As compensation I would like to recieve the cost of the ventillator in cash and the discount on my next day at your motel. The receipt for the ventillator is fortunately still in my posession.

Thank you for your response in advance, and I hope we can work out my compensation as soon as possible. I would not want to bad mouth your company because of faulty equipment.

Yours faithfully,

[real name]

Text B

Dear motel manager!

My name is [real name]. Last week I stayed at your hotel in New Orleans, and I had problems with the air conditioner in my room. The first day it worked well, but after that, it broke down quite often. When I say it broke down, I mean that it blow out hot air instead of cold. When it is 30 degrees outside a properly working AC is a must. Because I had a lot of thing to do while in New Orleans. I had no time to find another place to stay, so I had to fix the problem myself. Of my own money I bought a fan for the room, which I left there so the next customer don't have to buy one.

I would like some compensation for this. My idea is that you and the other workers check all of the air conditioners, and if they do not work properly, fix them. I don't want money or anything else! I want some kind of proof, so about this manner. Videos and picture of the fixing process! I hope I've made myself clear. This was one of the worst experiences I have never been to.

Thank you for listening,

[real name]

CVLA categorises the high-rated Text A as a C1 level text. The subscores are B2.2 for the verb per sentence, C2 for the average lexical difficulty, C2 for the measure of less frequent words over more frequent ones ("BperA"), and A2.1 for the readability measure.

The low-rated Text B is analysed into the A2.1 overall category. In terms of the four measures, verbs per sentence received a rating of B2.1, average lexical difficulty A2.2, BperA A2.1, and the readability variable a pre-A1 score.

163 texts collected at the May 2017 and April 2022 BLE rounds have been analysed with the help of the CVLA; see the results in Figure 1. It shows that only 23 per cent and 34 per cent of the texts (n=15 and n=33) are at or above the B2.2 level, respectively. In other words, the majority of the texts at both exam rounds are below the B2+ level. In the figure, a vertical line shows that cut-off point. While these results are indeed surprising, because of the yet small and partial sample I refrain from the validation-related interpretation of these results, for the time being.

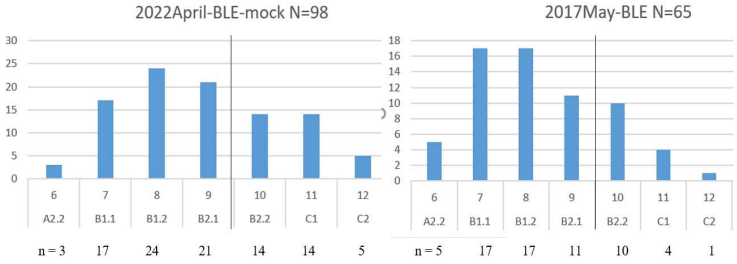

Figure 1. CVLA results of two BLE rounds, written text production. A vertical line shows the cut-off level.

Manual Analysis: An Example on Verb Tense/Aspect Distributions

As a second example, let me introduce one of our manual measures: the tense and aspect distribution in the verb phrases (VPs) of the written texts. I chose the same Texts A and B (see above), written on the same prompt (Topic 1 above) at the April 2022 BLE round, to illustrate the versatility difference in the verb tense/aspect distribution.

Table 3 shows the following information for each text: in the first row, the number of running words, the number of VPs, the number of finite VPs and within them the number of those beyond the Simple Present and Simple Past ("non-simple"), and the number of the non-finite VPs. The second row indicates the same per 100 words ("standardised"). In the third row those VPs are listed, according to verb tense/aspect category.

As Table 3 shows, the syntactic complexity characteristics of the two texts related to verb tense/aspect are quite different:

On the one hand, as far as the use of VPs other than the Simple Present and the Simple Past are concerned, the ratio between Text A and B is 3:2 (3.02 vs. 1.97). In other words, Text A uses one and a half times more other tenses, including the Present Progressive, the Present Perfect (in the passive), and modal VPs four times, while Text B uses the Present Perfect twice, one modal VP and the base form once.

On the other hand, the use of non-finite forms (VPs and clauses) also shows distinct frequencies, having a ratio of 2:1 (5.03 vs. 2.46). In other words, non-finite verb forms are twice as frequent in Text A than in Text B.

To summarise, we have two texts here that are dissimilar, among others, in their syntactic versatility: one uses various other tenses/aspects than the Simple Present and Simple Past, and also non-finite-verb forms including non-finite clauses, while the other one is more limited in those respects. At the moment I am in the process

of developing a quantitative index of syntactic versatility embracing these and other factors.

Table 3. Verb tense/aspect occurrences and their distribution in Texts A and B

Text A 199 words 29 VP, 19 finite (6 non-simple), 10 non-finite	**Text B** 203 words 33 VP, 28 finite (4 non-simple), 5 non-finite
Standardised 9.55 finite (3.02 non-simple), 5.03 non-finite	**Standardised** 13.80 finite (1.97 non-simple), 2.46 non-finite
Finite VPs 5 Simple Present: *feel, know, is, thank, hope* 8 Simple Past: *had, came about, was, made, had, had, was, had* 1 Present Progressive: *am writing* 1 Present Perfect Simple/passive voice: *have been misadvertised* 4 Modal VP: *would like, would like, can work out, would not want* **Non-finite** 8 VP: *to be compensated, to pay, to purchase, to leave, to take, to leave, to receive, to bad mouth* 2 Non-finite clause: *located, to relieve*	**Finite VPs** 13 Simple Present: *is, say, mean, *blow out, is, don't have, is, check, do not work, don't want, want, hope, thank* 11 Simple Past: *stayed, had, worked, broke down, broke down, had, had, had, bought, left, was* 2 Present Perfect Simple: *'ve made, *have been* 1 Modal VP: *would like* 1 Base form (imperative): *fix* **Non-finite** 5 VP: *working, to do, to find, to fix, to buy*

*=not Standard English; *non-simple*=other than Simple Present/Simple Past; *standardised*=per 100 words

A Method to Compare the Multivariate Systems

In the project we are working with over one hundred and ten variables in our attempt to find a limited, though not necessarily minimal, number of the most robust variables (and their co-occurrences) to predict a below-B2+ vs. B2+ level-specific patterning of syntactic complexity. Let me show how some statistical tests, more specifically correlation and factor analysis, are helpful to find those variables across the multivariate systems that corroborate each other. Data-coding and calculations are under way; the results below are partly based on Adamova (2022), Radnay (2017), Reményi and Velner (2022) and Velner (2022).

As the first example, the correlation matrix (Table 4) shows which variables in CVLA vs. some Lextutor variables vs. one of our own variables correlate with each other, on the basis of the April 2022 subcorpus data. Among the vertically presented CVLA variables, CVLAnum is the summary index before its

transformation into a CEFR-subcategory; the others are the ones introduced above: ARI is the readability variable, VperSent is the syntactic variable (verb per sentence), and AvrDiff and BperA are the lexical variables. Horizontally listed are, first, the Lextutor variables: the number of word families/types/tokens per text, followed by the 1,000 most frequent words on the token level (K1), the second most frequent 1,000 tokens (K2), all the tokens above those (K3up), followed by lexical density, i.e., the proportion of content words per all the words per text. The column on the right is our own variable, which calculates the proportion of K1 tokens compared to all the others (rest/K1 tokens). Shaded correlation coefficients are significant at the $p < 0.01$ and $p < 0.05$ levels.

Table 4. Spearman correlations between CVLA, Lextutor and one of our own variables (April 2022 subcorpus, N=98)

CVLA variables		Lextutor variables								Own
		Fami-lies	Types	Tokens	K1 tokens	K2 tokens	K3up tokens	Type/ token	Lex density	rest/K1 tokens
	CVLAnum	0.034	0.066	-0.084	-0.263**	0.471**	0.307**	0.233*	0.167	0.535**
	ARI	0.069	0.072	0.011	-0.093	0.404**	0.058	0.038	0.151	0.279**
	VperSent	-0.011	-0.008	0.039	0.048	0.102	-0.115	-0.130	-0.106	-0.069
	AvrDiff	0.002	0.004	-0.174	-0.396**	0.543**	0.444**	0.371**	0.249*	0.747**
	BperA	0.019	0.029	-0.167	-0.365**	0.407**	0.492**	0.328**	0.229*	0.663**

*** p < 0.01, * p < 0.05 (shaded)*

In the table, among the CVLA variables, VperSent is the only one that does not seem to correlate with the Lextutor and our own variables. This is no surprise, as VperSent is a syntactic complexity variable while the others are related to lexical complexity—although the two are related, in this subcorpus they seem not to be moving together. The rest of the CVLA variables are all corroborated by some of the other variables: for example, "AvrDiff" vs. "rest/K1 token" have a strong positive correlation (r=0.747), that is, the higher the average word difficulty value in the CVLA, the higher the ratio of above-K1 tokens in the texts, and vice versa. Similarly, "BperA" vs. "rest/K1 tokens" have a positive strong correlation between them (r=0.663), and "AvrDiff" vs. "K2 tokens" have a medium-strength positive correlation (r=0.543). The medium-level negative correlations between "K1 token" and "CVLAnum", "AvrDiff" and "BperA" mean that the higher the ratio of high-frequency K1 words tends to be in a text, the lower will be the value of the other variables. As far as the traditionally used "Family", "Type" and "Token" variables are concerned, they seem not to be corroborated by the CVLA variables.

Factor analysis is used as another way to detect common features among the almost ten dozen variables we are working with. This statistical test is usually employed to unearth the invisible, latent variables that lie beneath our surface variables, which are only surface manifestations of those hidden, latent ones. Table 5 shows the component matrix of the same variables used above (cf. Table 4), this time based on the May 2017 subcorpus. The table shows that Principal Component

Analysis (using Varimax rotation) yields five component factors that explain 84.17 per cent of the total variance (the rest of the factors are dropped due to their <1 eigenvalues).

Table 5. Factor analysis: Component matrix (CVLA, Lextutor and one of our own variables; May 2017 subcorpus)

	Component				
	1	2	3	4	5
Topic	0.165	-0.044	0.014	-0.002	0.962
Families	0.183	0.913	-0.017	0.183	-0.068
Types	0.176	0.938	0.002	0.118	0.015
Tokens	-0.034	0.928	0.045	-0.265	-0.011
K1 tokens	-0.310	0.875	0.015	-0.281	-0.022
K2 tokens	0.662	0.282	0.233	0.253	0.160
K3up tokens	0.832	0.192	-0.114	-0.221	-0.223
Type/token	-0.015	-0.111	-0.027	0.712	-0.024
Lex density	0.167	0.031	-0.030	0.726	0.035
rest/K1 tokens	0.929	-0.061	0.085	0.130	-0.032
CVLAnum	0.649	-0.065	0.672	0.042	0.129
ARI	0.215	0.070	0.916	0.066	0.039
VperSent	-0.082	-0.011	0.944	-0.158	-0.050
AvrDiff	0.894	-0.096	0.105	0.154	0.234
BperA	0.884	-0.018	0.083	0.067	0.188

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

To figure out what those five latent variables are, let us see which correlate strongly with which of the surface variables (the strongest correlations are highlighted for each). Table 5 shows that Factor 1 has the highest correlation with our own "rest/K1 token" variable, and also the "AvrDiff" and "BperA" variables (CVLA) and the "K2 tokens" and "K3up tokens" variables by Lextutor. All these point in one direction: this factor is about word frequency. In a similar fashion, Factor 2 is related to vocabulary breadth ("Families", "Types", "Tokens", "K1 tokens"). Factor 3 correlates strongly with the syntax-related CVLA variables "VperSent" and "ARI". Factor 4 correlates strongly with two traditional corpus linguistic measures: type-token ratio and lexical density. As a shortcut, I named

this factor "Tom Cobb's sweethearts" (see Cobb 2023). Factor 5 is correlated only with the topic of the formal letter chosen at the BLE round, which confirms the validity of this analysis: the topic choice at the exam did not affect any of these variables of syntactic and lexical complexity.

Our data-coding and calculations are still underway. Still, the above partial investigations already seem to show that the way to cope with the amount of data based on the numerous variables yielded by the multivariate analytical systems we are working with is finding a limited, though not necessarily minimal, number of variables that will predict the patterns of syntactic and lexical complexity characteristics of B2+ vs. below-B2+ L2-English writing. And more generally, cross-examining those variables through correlation analysis and factor analysis can provide a grasp to handle the amount of data that is characteristic of syntactic and lexical analyses across all L2-English proficiency levels.

Conclusion

In this paper a research project has been introduced that seeks to find solutions to two types of problems. One of them is local and practical: a language exam is to be validated through an externally supported quantitative analysis, based on syntactic and lexical complexity. The other problem is global and both theoretical and practical: L2-English proficiency levels are to be modelled, which in our case is based on characteristic patterns of syntactic and lexical complexity.

To find answers to both problems in a short paper is, obviously, impossible. What I am suggesting here (which is an answer to my present research question) is that there is indeed a quantitative method of analysis that can help us pick the most relevant features of syntactic and lexical complexity. A statistical meta-analysis of the variables related to the multivariate analytical systems is recommended, to select the variables that corroborate each other within and across those multivariate systems. The ultimate variables thus proven to characterise L2-English proficiency at various levels may inform both the teaching-learning process, and provide the basis for partial automated evaluation of those texts at language exams, parallelly to human rating – to relieve trained human raters, and also to increase the reliability of those exams.

A long-term aim of the language testing profession internationally is to move towards automated language assessment as far as the quantifiable features of language proficiency are concerned. But the statistical road described above is only one of the possibilities ahead. The other road, now emerging, is based on large language models and supervised machine learning (Lu and Bluemel, 2021).

References

Adamova, Karlygash 2022. *A Multidimensional Analysis of L2 Learners' Writings at the B2+ Level.* Unpublished manuscript. Budapest: Pázmány Péter Catholic University.

Ai, Haiyang 2022. *Web-Based L2 Syntactical Complexity Analyzer*. Online/downloadable computer programme. https://aihaiyang.com/software/l2sca/

Ai, Haiyang, and Xiaofei Lu 2013. "A Corpus-Based Comparison of Syntactic Complexity in NNS and NS University Students' Writing". In *Automatic Treatment and Analysis of Learner Corpus Data*, edited by Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, 249-64. Amsterdam: Benjamins.

BAAL 2021. *Recommendations on Good Practice in Applied Linguistics.* British Association for Applied Linguistics. https://www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf

Biber, Douglas, Bethany Gray, and Shelley Staples 2016. "Predicting Patterns of Grammatical Complexity across Language Exam Task Types and Proficiency Levels". *Applied Linguistics* 37 (5): 639-668.

Bulté, Bram, and Alex Housen 2012. "Defining and Operationalising L2 Complexity". In *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA,* edited by Alex Housen, Folkert Kuiken, and Ineke Vedder, 21-46. Amsterdam: Benjamins.

——— 2014. "Conceptualizing and Measuring Short-Term Changes in L2 Writing Complexity". *Journal of Second Language Writing* 26: 42-65.

CEFR 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg: Council of Europe – Cambridge: Cambridge University Press.

——— 2018. *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume with new descriptors.* Strasbourg: Council of Europe.

Cobb, Tom 2023. *LexTutor Vocabprofiler*. Online computer programme. https://www.lextutor.ca/vp/

Crossley, Scott, and Danielle McNamara 2014. "Does Writing Development Equal Writing Quality? A Computational Investigation of Syntactic Complexity in L2 Learners". *Journal of Second Language Writing* 26: 66-79.

CVLA 2023. *CVLA: CEFR-Based Vocabulary Level Analyzer (ver. 2.0)*. Online computer programme. https://cvla.langedu.jp/index.html

Housen, Alex, and Folkert Kuiken 2009. "Complexity, Accuracy and Fluency in Second Language Acquisition". *Applied Linguistics* 30 (4): 461-73.

Lei, Lei, Ju Wen, and Xiaohu Yang 2023. "A Large-Scale Longitudinal Study of Syntactic Complexity Development in EFL Writing: A Mixed-Effects Model Approach". *Journal of Second Language Writing* 59: 100962. DOI: 10.1016/j.jslw.2022.100962

Lu, Xiaofei 2010. “Automatic Analysis of Syntactic Complexity in Second Language Writing”. *International Journal of Corpus Linguistics* 15 (4): 474-96.

——— 2017. “Automated Measurement of Syntactic Complexity in Corpus-Based L2 Writing Research and Implications for Writing Assessment”. *Language Testing* 34 (4): 493-511.

Lu, Xiaofei, and Brody Bluemel 2021. “Automatic Assessment of Language”. In *The Cambridge Introduction to Applied Linguistics,* edited by Susan Conrad, Alissa Hartig, and Lynn Santelmann, 86-98. Cambridge: Cambridge University Press.

Mancilla, Rae, Nihat Polat, and Ahmet Akcay 2015. “An Investigation of Native and Nonnative English Speakers’ Levels of Written Syntactic Complexity in Asynchronous Online Discussions”. *Applied Linguistics* 38 (1): 1-24.

Nini, Andrea 2019. “The Multidimensional Analysis Tagger”. In *Multidimensional Analysis: Research Methods and Current Issues,* edited by Tony Berber Sardinha, and Marcia Veirano Pinto, 67-94. London: Bloomsbury Academic.

——— 2021. *The Multidimensional Analysis Tagger.* Online/downloadable computer programme. https://sites.google.com/site/multidimensionaltagger

Ortega, Lourdes 2003. “Syntactic Complexity Measures and Their Relationship to L2 Proficiency: A Research Synthesis of College-Level L2 Writing”. *Applied Linguistics* 24 (4): 492-518.

Polat, Nihat, Laura Mahalingappa, and Rae Mancilla 2019. “Longitudinal Growth Trajectories of Written Syntactic Complexity: The Case of Turkish Learners in an Intensive English Program”. *Applied Linguistics* 41 (5): 688-711.

PPCU 2017. *Basic Language Exam: Writing Criteria*. Internal manuscript. Budapest: Pázmány Péter Catholic University.

——— 2022. *Mock Exam: Composition Writing – Formal Letter. Task Description.* Internal manuscript Budapest: Pázmány Péter Catholic University.

Radnay, Zsanna 2017. *Corpus-Based Validation of the Basic Language Examination at Pázmány Péter Catholic University: An Analysis of the Written Production.* Unpublished BA thesis. Budapest: Pázmány Péter Catholic University.

Reményi, Andrea Á., and Patrik Velner 2022. *Validating an EFL Examination through the Manual and Automated Quantitative Analysis of Candidates’ Written Texts.* Talk at the ESSE2022 conference, Mainz, Germany.

Taguchi, Naoko, William Crawford, and Danielle Wetzel 2013. “What Linguistic Features are Indicative of Writing Quality? A Case of Argumentative Essays in a College Composition Program”. *TESOL Quarterly* 47 (2): 420-30.

Uchida, Satoru, and Masashi Negishi 2018. “Assigning CEFR-J Levels to English Texts Based on Textual Features”. In *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018),* edited by Yukio Tono and Hitoshi Isahara, 463-67. Takamatsu: APCLA.

Van Ek, Jan Ate, 1975. *The Threshold Level.* Strasbourg: Council of Europe.

Van Ek, Jan Ate, and John L. M. Trim 1991a. *Threshold 1990*. Strasbourg: Council of Europe – Cambridge: Cambridge University Press.
——— 1991b. *Waystage 1990*. Strasbourg: Council of Europe – Cambridge: Cambridge University Press.
——— 2001. *Vantage*. Strasbourg: Council of Europe – Cambridge University Press.
Velner, Patrik 2022. *CEFR-Based Assessment with Computational Methods: Investigating Accuracy and Complexity in English Learner Texts*. Unpublished MA thesis. Budapest: Pázmány Péter Catholic University.