Syntactic and lexical complexity in L2 English academic writing

O'Leary, John A.; Steinkrauss, Rasmus

Published in:
Ampersand

DOI:
10.1016/j.amper.2022.100096

Ampersand 9 (2022) 100096

Contents lists available at ScienceDirect

Ampersand

journal homepage: www.elsevier.com/locate/amper

Syntactic and lexical complexity in L2 English academic writing: Development and competition

John A. O'Leary [1], Rasmus Steinkrauss [*]

Faculty of Arts, Department of Applied Linguistics, University of Groningen, the Netherlands[1]

ARTICLE INFO

Keywords:
Academic writing
Lexical complexity
Syntactic complexity
Dynamic systems theory
Second language development
Complexity measures

ABSTRACT

Academic writing is both challenging and essential for university students studying in a second language. Linguistic complexity is a useful indicator of second language development, including academic writing. One area of interest is the relationship between lexical and syntactic complexity, which some studies have shown to be in competition in learner writing. However, given the dynamic nature of language development, identifying reliable complexity measures is not easy. This study aims to identify which measures of linguistic complexity best capture writing quality and what kind of interaction exists between lexical and syntactic complexity by analysing 42 L2 English essays of three Dutch undergraduates using 12 complexity measures. A correlation analysis using pairwise ratings revealed that the lexical frequency profile and complex nominals per sentence best predicted academic quality. Negative correlations between lexical diversity and phrasal complexity measures suggest a competitive relationship between lexical and syntactic complexity.

1. Introduction

Second language (L2) English academic writing is an interesting area of focus not only because it is a highly valued skill necessary for academic qualifications across the world, but also because it is a highly challenging form of language production which tests a learner to the limit of their abilities. As such, analysing academic writing provides an insight into the sort of linguistic development taking place in advanced level learners. The present study focuses on linguistic complexity in the academic writing of three advanced L2 learners of English produced over a period of three to four years. Linguistic complexity is a useful construct to focus on in academic writing since it can be regarded as a reliable index of language development and progress (Verspoor et al., 2017, p. 1) and has often been used as such (Norris and Ortega, 2009; Bulté and Housen, 2012).

For the present study, 12 complexity measures will be applied to texts from three Dutch university students written over a period of three to four years. In particular, the study is interested in which complexity measures best capture the overall quality of academic writing and what kind of interaction exists between lexical and syntactic complexity in the three advanced L2 learners.

In this study, we will take a dynamic perspective on linguistic development and view language as a complex, dynamic system consisting of several interacting subsystems (de Bot et al., 2007). The dynamic view on language has several consequences for what kind of development may be observed in learners. Firstly, due to the different subsystems interacting with each other and with the environment, linguistic development is expected to be non-linear. Secondly, due to the different resources available to each individual learner, the developmental trajectories will be different for every learner. And thirdly, since resources are limited, they have to be distributed among the different interconnected subsystems, which may result in (temporary) competition between subsystems and one subsystem developing at the expense of another.

For the study of linguistic complexity, this raises the question of whether it is possible to identify valid measures of complexity which, in spite of the individual, non-linear development and the involvement of many different subsystems in language development, can reliably capture the quality of a learner's academic writing. Additionally, the question arises in what way the development of different aspects of complexity such as grammatical and lexical complexity interact over time. These questions will be the focus of the present study. To a certain extent, this study mirrors that of Verspoor et al. (2017) which analysed which complexity measures captured both the quality of academic

* Corresponding author.
E-mail addresses: johnol22@hotmail.com (J.A. O'Leary), r.g.a.steinkrauss@rug.nl (R. Steinkrauss).
[1] Present address: International Business College, Dongbei University of Finance and Economics, Dalian, China.

https://doi.org/10.1016/j.amper.2022.100096
Received 15 April 2022; Received in revised form 6 October 2022; Accepted 7 October 2022
Available online 1 December 2022

writing and development over time. However, this study differs from Verspoor et al. in several respects, including the choice of which complexity measures to use and in focussing on the interaction between subsystems rather than the development of individual subsystems over time.

In the first part of the study, a hitherto little-used system of pairwise comparison will be used to rate the quality of all the texts written by one of the learners. These ratings will be correlated with the 12 complexity measures in order to determine which ones best capture the overall quality of this learner's academic writing. In the second part, correlations between measures associated with the lexical and syntactic subsystems will be analysed to explore any potential interactions.

2. Background

2.1. Academic writing

Present-day academic writing is arguably a distinctive register (Biber and Gray, 2010) which is highly challenging for L2 learners in terms of production. It is also a complex and elaborate form of language, but as work by Biber and Gray (2010) has revealed, not always in ways conventionally supposed. For them, English academic writing is not characterised by the use of longer sentences and more elaborate clausal subordination (when compared to speech), but rather is more structurally 'compressed' with phrasal (non-clausal) modifiers embedded in noun phrases. For example, while subordinate clauses such as complement clauses and adverbial clauses are more frequent in conversation, phrasal modifiers such as adjectives or nouns modifying head nouns are more prevalent in English academic writing (p. 7). On this view, complexity in academic writing (as compared to spoken language) is to be found more in its use of embedded phrases than in its use of clausal subordination.

On this basis, Biber et al. (2011) proposed that the use of embedded phrases in L2 student writing emerges in several hypothesized stages, each more elaborated than the last. This has been partially corroborated by Parkinson and Musgrave (2014) who found that the use of noun modifiers in L2 academic writing increased with proficiency and by Crossley and McNamara (2014) who found that L2 writers produced texts that contained more and more nouns and phrasal complexity over the course of a semester.

In addition to the structural features identified by Biber and Gray (2010), academic language also has a distinctive lexical character. Work by Coxhead (2000) sought to identify words outside the General Service List (West, 1953) which occurred frequently across a range of academic texts. What resulted was the Academic Word List, a list of 570 word families which account for 10.0% of the tokens in academic texts (Coxhead, 2000, p. 222), considerably more than the 1.4% found in fiction writing for example (p. 225). According to Coxhead's research, the Academic Word List and General Service List combined account for 86.1% of tokens found in academic texts while Douglas (2013) found the two lists covered 94% of a typical paper produced by an entry-level (native English speaking) university student.

There is good reason to believe that a vocabulary profile of learner writing is a strong indicator of development. Morris and Cobb (2004) demonstrated that students' use of words from the Academic Word List correlated significantly with grades achieved, while work by Verspoor et al. (2017) showed that academic word usage is linked to independent ratings of text quality. Given the distinctive lexical character of academic writing, these studies suggest that attention to the vocabulary used at various stages of development could provide a useful means of both capturing the quality of student writing and tracing development over time. The current study therefore focuses on syntactic and lexical complexity in English academic writing, and it does so from a Dynamic Systems Theory perspective.

2.2. Dynamic Systems Theory

Dynamic Systems Theory (DST) is applied in the field of language studies as a tool to understand language development which it views as a kind of complex system. This complex system is composed of many different interdependent subsystems including the lexicon, syntax, morphology, phonology and pragmatics (Larsen-Freeman, 1997, p. 149). The characteristics of a complex system according to this view are numerous, but those germane to the present study are that language development is non-linear, involves complete interconnectedness of the system and chaotic variation over time (de Bot and Larsen-Freeman, 2011).

For language development to take place, there must be resources to keep the process of growth going (de Bot et al., 2007, p. 11). These include internal resources such as the capacity to learn, conceptual knowledge and motivational resources, and external resources such as the language used in the environment and learning materials (p. 11). The development itself is proportional to (amongst other things) the available resources (van Geert, 2008, p. 190) and, since these resources are finite, they have to be distributed amongst different subsystems (such as the lexicon and syntax) often resulting in competition between those sub-systems (Verspoor et al., 2017, p. 2).

2.3. Competition

The kind of competition between subsystems predicted by DST might be portrayed as a Trade-off Hypothesis (Skehan, 2009) which predicts that a greater concentration of resources into the development of one subsystem might, all else being equal, result in lower performance in others. Indeed, previous research within a complexity-accuracy-fluency framework has indicated that tasks often lead to development within two of these performance areas, but rarely all three (p. 512). For example, in a study testing the effects of task type on speaking performance, it was found that in the case of non-native speakers, there was a negative correlation between lexical sophistication and structural complexity. Skehan concludes that due to problems occurring through the demands of lexical retrieval, the syntax of non-native speakers is, in some sense, derailed (p. 516).

A case study of learner academic writing by Verspoor et al. (2008) further investigated the interaction between lexical and syntactic complexity. The study found an interesting relationship between type-token ratio, a lexical diversity measure, and average sentence length, a syntactic measure. When the two measures were traced over 18 assignments written by the learner, a competitive interaction emerged. When the type-token ratio increased, average sentence length decreased, with the reverse also being true. When represented graphically, an oscillating pattern emerged as the two measures interacted producing 'waves' which appeared to alternate almost perfectly (p. 223). This lends support to the idea of a trade-off effect between different aspects of L2 development and suggests that a similar competition and oscillating patterns between lexical and syntactic complexity might be observed in the present study.

2.4. Defining linguistic complexity

The notion of 'complexity' was alluded to in previous sections without proper elucidation of its meaning. Indeed, Bulté and Housen (2012) complain that many L2 studies inadequately define or fail to define altogether this term resulting in mixed and sometimes contradictory results. They themselves describe language complexity as "the number of discrete components that a language feature or a language system consists of, and as the number of connections between the different components" (p. 24) and identify linguistic complexity as one component of this (along with both propositional and discourse-interactional complexity). However, the authors make clear that this definition falls under an 'absolute approach' which defines

complexity in objective quantitative terms as the number of components and connections within a language system. This contrasts with a 'relative approach' which sees a language feature as complex if it presents a challenge to the language user to internalise or process it. In this relative sense, the notion of complexity can be understood as 'difficulty'.

Pallotti (2015), who also recognises a distinction between an absolute, 'structural' approach and a relative, 'developmental' approach, advocates a 'simple view' of complexity in which its use is limited to only this kind of structural complexity. However, while freeing complexity from any theoretical assumptions about development would allow for a more consistent application across studies, it also presents serious problems where frequency measures are concerned. It is difficult to see, for example, how a lexical frequency measure can capture structural complexity since a less frequent word used once in a text does not differ in terms of number or relational parts to a more frequent word used once in a text. Likewise, a less frequent grammatical feature such as the use of embedded phrases in academic writing may not count as more complex than the use of clausal subordination common in spoken conversation.

One could argue, of course, that in this case frequency measures should not be considered as complexity measures at all, but as something else entirely. However, on further reflection, it seems that frequency measures *do* capture complexity of an absolute kind, but in the language system of the individual learner rather than in the written or spoken language they produce (Bulté and Housen, 2012, p. 26). If a low-frequency vocabulary item or grammatical form is found in a sample of writing, it is probable that complexity exists in the language system of the individual who produced the text, likely in the form of a larger vocabulary or a greater repertoire of grammatical structures.

There is limited space here to fully develop an answer to these problems, but in light of the issues outlined above, the present paper will adopt a softer view of complexity than the simple view advocated by Pallotti. Complexity will indeed be considered as related to the number of elements of a linguistic item and their relational patterns, but less importance will be attached to whether that complexity manifests itself on the concrete or abstract level. A frequency measure is, on this view, a complexity measure if it can sensibly be considered an index of the vocabulary size or grammatical repertoire of an L2 learner, even if the language produced cannot itself be considered 'complex'.

Bulté and Housen (2012) further propose that linguistic complexity be divided into two principal sub-components, *lexical* complexity and *grammatical* complexity, with the latter analysable into *syntactic* and *morphological* complexity. These are useful distinctions and will be adopted in the present study, although morphological complexity will be ignored on the grounds that its development is most evident in the early stages of L2 development rather than in the academic writing of advanced learners (at least in English). Bulté and Housen further propose that lexical complexity is manifested at the observational level in L2 performance in terms of *density*, *diversity* and *sophistication* of L2 lexical items and collocations and, based on work by Norris and Ortega (2009), propose that syntactic complexity is manifested as *sentence* (or *sentential*) complexity, *clausal* complexity and *phrasal* complexity. This framework will be adopted in this study. The following section will describe how each of the six constructs will be operationalised, wherein each construct is operationalised with two different measures. Due to space constraints, it is only possible to provide a brief overview of each measure.

2.5. Operationalising lexical complexity

2.5.1. Lexical sophistication

Lexical sophistication will be measured with the *Beyond 2000* and the *P_Lex* measures. Beyond 2000 is derived from the lexical frequency profile (LFP) proposed by Laufer and Nation (1995) and involves taking a text as raw input and describing its lexical content in terms of frequency bands. The 'classic' version of the LFP analysed texts into four bands: the first 1000 most frequent words, the second 1000 most frequent, the Academic Word List (Coxhead, 2000) and finally 'off list' words which do not occur in any of the lists (Laufer, 2012, p. 1). To deal with the LFP statistically, Laufer (1995) recommended Beyond 2000, a condensed profile of the LFP calculated from the percentage of words in the text which are not within the first 2000 words. Laufer argues that this provides a reliable and valid measure which has the advantage of producing a single figure.

However the LFP is calculated, a significant shortcoming is that the percentage of words falling outside the first two categories is typically very small (rarely exceeding 10%), thus necessitating a longer sample length to achieve stable measures. In an attempt to address this issue, Meara and Bell (2001) proposed P_Lex, a measure which works by looking at the distribution of difficult words in a text and calculating an index which reflects the likelihood of these words occurring. This involves dividing the text into 10-word segments and counting the number of difficult words which occur within each segment. Difficult words are those which, except for proper nouns, numbers and geographical derivatives, fall outside the 1000 most frequent words. This data is then transformed into a *lambda* figure, typically ranging from 0 to 4.5, which has the advantage of being less sensitive to text length than the LFP (Meara and Bell, 2001).

2.5.2. Lexical diversity

Lexical diversity is assessed through the *Guiraud* and *Vocd-D* measures. Both the Guiraud (1954) and Vocd-D (McKee et al., 2000) (based on D; Malvern and Richards, 1997) are derived from the type-token ratio (TTR) measure (Templin, 1957). TTR, however, is sensitive to sample size (Malvern and Richards, 1997, p. 59) and both the Guiraud and Vocd-D were designed to mitigate this effect through mathematical transformations of the TTR. The Guiraud attempts this by dividing the number of types by the square root of the number of tokens while Vocd-D uses a random sampling technique to calculate an average TTR score for the whole sample. Studies such as Zenker and Kyle (2021) have cast doubt on both measures' independence of text length, particularly for texts of less than 200 words. However, given that the present study uses samples of roughly the same length (ranging from 270 to 330 words), this is not a significant issue.

2.5.3. Lexical density

The final sub-component of lexical complexity proposed by Bulté and Housen (2012) and targeted in this study is lexical density. The term 'lexical density' was coined by Ure (1971) to describe a measure of the relationship between the number of words with lexical properties (as opposed to grammatical properties) as a percentage of the total number of words in a text (O'Loughlin, 1995, p. 221). Halliday (1985) proposed a revision to Ure's method by recommending that the number of lexical items be calculated as a percentage, not of the total number of words, but rather the total number of clauses. In this study, both Ure's and Halliday's method will be used to operationalize lexical density.

2.6. Operationalising syntactic complexity

2.6.1. Sentential complexity

Sentential complexity will be operationalised as both the average number of words and the average number of morphemes per sentence. Following Norris and Ortega (2009), average sentence length in both forms will be considered a "global or generic metric of linguistic complexity" since it is impossible to establish on the basis of this measure whether any increase (or decrease) in sentence length is due to a change in clause or phrase length, or a combination of the two (p. 561).

2.6.2. Clausal complexity

Following Norris and Ortega (2009), clausal complexity will be considered as measurable by "any metric with clause (or subordinate or dependent clause) in the numerator" (p. 561). A clause will be

considered as "a structure with a subject and a finite verb (a verb with a tense marker)" based on Hunt (1965, p. 15). Bulté and Housen (2012) consider both C/S (clauses per sentence) and DC/C (dependent clauses per clause) measures as suitable measures of clausal complexity and, in their own survey of complexity measurement used in the academic literature, identified the two measures as the most frequently used to explore this aspect of syntactic complexity (p. 30). The current study will also use these measures.

2.6.3. Phrasal complexity

While sentence length must be considered a global metric of linguistic complexity for the reasons outlined above, clause length is unaffected by the amount of subordination in production and so taps into complexification sub-clausally, at the phrasal level (Norris and Ortega, 2009, p. 561). An increase in clause length can only result from phrasal elaboration (via adjectives, adverbs, prepositional phrases, or non-finite clauses) or the use of nominalisations (p. 561) and so average clause length (ACL) can be considered an index of phrasal complexity.

Bulté and Housen (2012) observe that phrasal complexity has, to a large extent, been neglected in the literature as a sub-component of linguistic complexity (p. 29). It is difficult, therefore, to identify a tried and tested measure to tap into this construct. One option is to investigate the frequency of the complex nominal (a grouping of words which together function as a noun) which may increase as academic writing develops. Ravid and Zilberbuch (2003) found that texts produced in Hebrew by both school-age and adult writers produced more complex and diverse nominals the older and more experienced the writers were. Given Biber and Gray's (2010) observations that phrasal elaboration is a key characteristic of academic writing, measuring the number of complex nominals per sentence (CN/S) may offer a valuable insight into development at this stage.

2.7. Research questions

The present study addresses weaknesses in previous studies such as Verspoor et al. (2017) relating to sample size and selection. The sample size of 200 words used by Verspoor et al. was probably insufficient, particularly when measuring whole sentence lengths or academic words which occur infrequently over this length. It was also felt that, given the uneven nature of a typical piece of academic writing, Verspoor et al.'s method of random sample selection would likely lead to unrepresentative samples. The present study, therefore, uses larger sample sizes and a more robust method of sample selection to mitigate these issues. The study also focusses on competition over time, adding to the growing body of work investigating the interaction between different aspects of linguistic complexity longitudinally (for an overview, see Bulté and Housen, 2020), i.e. outside of a (often cross-sectional) focus on the effect of task on complexity, accuracy, and fluency (e.g., Alexopoulou et al., 2017).

The main research questions of this study concern the interpretation of complexity measures, both lexical and syntactic, when applied to a longitudinal corpus of academic writings of three advanced L2 English learners. With regard to this, two research questions were formulated:

1. Can measures of linguistic complexity capture the overall quality of the academic writing of an advanced L2 learner of English, and if so, which measures are they?
2. What is the relationship between the lexical and syntactic subsystems of the three learners?

The first question will be operationalised as which single measures of linguistic complexity taken from the academic essays of a single learner correlate most strongly with ratings of these essays provided by a group of human judges. Given the nature of academic writing as outlined above, it is predicted that the measures related to phrasal elaboration and lexical sophistication will be most strongly tied to text ratings.

The basic operationalisation of the second question will be whether or not a negative correlation exists between the lexical and syntactic measures taken from each learner. The idea is that a negative correlation signals that two measures are competing for resources. On the basis of the Skehan (2009) and Verspoor et al. (2008) findings, it is predicted that a competitive relationship will be found between at least some of the lexical and syntactic measures. It is difficult, however, to predict exactly which measures will interact and what the nature of these interactions will be. In view of DST, variation across the three learners is anticipated, but common patterns may also emerge which could form the basis of future research.

3. Method

3.1. Participants and essays

All the texts analysed as part of the study were taken from three Dutch female university students. They all completed English Language and Culture degrees at Dutch universities (students A and B at the University of Groningen, Student C at Radboud University) and were aged 18 to 21 during the time the texts were written. They can all be considered advanced level students since both programmes stipulated an IELTS score of 6.0 or more (6.5 in the case of Groningen) or an equivalent to gain admission. The essays were all written during the course of their study and were well distributed over this time. The average gap between essays was 3.2 months (SD = 2.3). The longest interval between two essays was 11 months with the second longest 7 months. The essays were all written on topics related to the degree programme and focussed on either literature or linguistics. All the essays were written without a time limit with free access to dictionaries and other resources.

In total, the three students consented to 56 essays being analysed, 42 of which were deemed suitable. Reasons for rejection included that the essays were too short for analysis, had multiple authors, were written in a question and answer format which may have influenced writing style, or contained significant overlap with a previous essay (some redrafts were submitted). Table 1 below provides details of the final 42 essays selected for analysis.

3.2. Sample selection

Once the 42 texts had been chosen, they were divided by student and numbers were assigned to each text according to the order they were written. A representative sample from each text was then extracted. In line with Verspoor et al. (2017), introductions and abstracts were removed along with any direct quotes or references. Whilst Verspoor et al. opted for a 200-word sample, a larger 300-word sample was selected to improve reliability. To avoid selecting unrepresentative samples from unevenly written texts, the current study avoided Verspoor et al.'s method of randomly selecting samples. Instead, the method adopted involved extracting from each text every possible 300-word sample of continuous text (within a margin of 10% to preserve sentence integrity) and then comparing each extracted sample with the overall text to find the closest match in terms of complexity.

The sample extraction was performed using a bespoke program

Table 1
Summary of the analysed essays.

	Student A	Student B	Student C
Number of essays	13	17	12
Date of first essay	Jan 2012	Oct 2014	Oct 2014
Date of last essay	Dec 2015	Jun 2017	Aug 2018
Total words	11730	31753	28897
Average length (words)	902 (SD = 409)	1868 (SD = 1646)	2408 (SD = 2630)

written in Python (O'Leary, 2022). The program takes the first sentence of a text as a starting point and then iterates through subsequent sentences whilst taking a cumulative word count. When 300 words are reached (or as close as possible without violating sentence integrity), the sentences to this point are taken as the first sample. The process is then repeated using the second sentence of the text as a starting point to produce the second sample. The program then proceeds through the remainder of the text in the same manner.

From each sample, an average sentence length in words (ASL) and average word length in letters (AWL) measurement was taken and compared to the same measurements taken from the original text using Z-tests. Z-tests were chosen in preference to t-tests since the population variance was known (both average sentence and word lengths for the original text had been established). These Z-test scores were squared (to remove negative values) and added to produce a single overall value for each sample. All the samples were then ranked according to this value and the sample with the smallest overall difference from the original text was selected as the sample to be analysed. The rationale behind using ASL and AWL was that, firstly, the two measures are simple to calculate and require minimal preparation and, secondly, they both have a long history as reliable measures of complexity (Norris and Ortega, 2009; Bulté and Housen, 2012); for example, the Flesch–Kincaid readability tests (Flesch, 1948) use both measures. It seems reasonable to assume that a sample containing a similar degree of complexity to the whole text will be most representative sample of that text.

To illustrate the method, Table 2 provides a summary of the output for text 16, student B. From this text of 6036 words, 206 300-word samples were extracted (the number of possible samples varies from text to text), with sample number 119 proving to be the most similar to the original text in terms of AWL and ASL.

3.3. Complexity measures

Before discussing how the different complexity measures are being calculated, Table 3 lists them together with some studies that have used the same measures.

3.3.1. Lexical complexity

Beyond 2000 was calculated using AntWordProfiler (Anthony, 2014) which produces an LFP from which Beyond 2000 can be derived. The package includes the New General Service List (Browne et al., 2013) which serves to represent the first and second bands of most frequent words along with the Academic Word List compiled by Coxhead (2000). Proper nouns were removed before the off list was calculated in line with Laufer and Nation (1995).

The P_Lex lambda value was calculated using online software provided by Meara (2018) which uses a vocabulary list developed by Nation (1984). The program identifies 'difficult' words in a text which the user can manually reclassify as 'easy' where appropriate. In line with Meara and Bell (2001), proper nouns, numbers and geographical derivatives were reclassified as 'easy'.

The Vocd-D scores were calculated using the function *vocd* (McKee et al., 2000) in CLAN (MacWhinney and Snow, 1990). The Guiraud score was derived by applying the appropriate calculation to the type and token values produced by *vocd*.

Both Ure and Halliday's density measures were calculated using a combination of CLAN and the L2 Syntactical Complexity Analyzer (L2SCA) developed by Lu (2010) and implemented in TAASSC (Kyle, 2016). The total number of clauses was calculated using L2SCA (the program provides an average clauses per sentence figure which can be multiplied by the number of sentences to produce this figure) while the number of lexical words was calculated relying on the part-of-speech tagging by CLAN. Following O'Loughlin's (1995) taxonomy, all nouns (including proper nouns), adjectives and verbs (including copulas and participles) were counted as lexical items, as were those adverbs which could be considered "adverbs of time, manner and place" (p. 228). This distinction was made manually based on a list of all adverbs produced by CLAN from the texts. In deviation from O'Loughlin's taxonomy, the verbs 'to be' and 'to have' were treated as lexical except when they occurred as auxiliary verbs. This is in line with the common practice of classifying non-auxiliary verbs as lexical (for example, Huddleston and Pullum, 2005, p. 18).

3.3.2. Syntactic complexity

Both the average sentence length in words and in morphemes were calculated using the figures provided by CLAN's morphosyntactic analysis tool *MOR*, while the measures for clausal and phrasal complexity were obtained using L2SCA. The C/S, DC/C and ACL measures were produced directly by the program, while the CN/S score was calculated by multiplying the C/S measure with the 'complex nominals per clause' measure provided by L2SCA.

3.4. Assessing text quality

To assess text quality, a pairwise comparison test was performed on the texts provided by student C. Since this procedure is not widely used in L2 development studies, it will be elaborated a bit more in detail. The design and principle behind the pairwise comparison test are based on the work of Thurstone (1927) and is a means of generating a quality rating by presenting a judge with a pair of objects (texts in this case) and asking them to state which one possesses a specified attribute to the greatest degree. Repeated comparisons of different pairs across judges allow the construction of a psychological scale of 'perceived quality' (Bramley, 2007, p. 246). Each text's location on this scale is determined by the number of comparisons 'won' or 'lost'.

A significant advantage of pairwise comparison is its simplicity. Judges are not required to rate texts in absolute terms according to a predetermined scale, but in terms only of their relative merit within a pair. This allows for the removal of the individual severities of the judges (Heldsinger and Humphry, 2010) which is especially important when non-expert judges are used (p. 247). Work by Heldsinger and Humphry (2010) demonstrated that pairwise comparison of writing scripts

Table 2
Samples extracted from text 16, student B.

Rank	Sample Number	Words	AWL	Z^2 (AWL)	ASL	Z^2 (ASL)	Combined Z^2 scores
–	Original	6036	5.16	–	28.34	–	–
1	119	311	5.15	0.0051	28.27	0.0003	0.0054
2	117	303	5.17	0.0015	27.55	0.0485	0.0499
3	113	278	5.19	0.0383	27.80	0.0203	0.0586
4	50	290	5.19	0.0375	29.00	0.0307	0.0682
…							
203	92	306	5.49	2.7597	38.25	5.5086	9.9402
204	149	309	4.72	2.4501	23.77	1.9020	10.2132
205	91	311	5.51	2.7592	38.88	6.2252	11.1505
206	93	291	5.45	2.7508	41.57	8.5910	11.7728

Note. In this case, sample number 119 was considered the most representative sample since its combined *Z*-test score was lower than any other sample.

Table 3
Measures used to operationalize complexity.

Construct	Measure	Abbrev.	Previous studies
Lexical Sophistication	Beyond 2000		Lemmouh (2008); Douglas (2013); Higginbotham and Reid (2019)
	P_Lex		Meara and Bell (2001); Skehan (2009)
Lexical Diversity	Guiraud		Yixin and Daller (2014)
	Vocd-D		Gebril and Plakans (2016)
Lexical Density	Ure's Method		Ishikawa (2007); To et al. (2013)
	Halliday's Method		Ishikawa (2007); To et al. (2013)
Sentential	Average Sentence Length (words)	ASL-w	Ishikawa (2007); Storch and Wigglesworth (2007)
	Average Sentence Length (morphemes)	ASL-m	Klee and Fitzgerald* (1985)
Clausal	Clauses per Sentence	C/S	Ishikawa (2007); Sercu et al. (2006)
	Dependent Clauses per Clause	DC/C	Ishikawa (2007); Sercu et al. (2006)
Phrasal	Average Clause Length	ACL	Ishikawa (2007)
	Complex Nominals per Sentence	CN/S	Ravid and Zilberbuch (2003)

*Study focussed on child language.

performed by teachers produced a scale which correlated highly with those produced from the results of a large-scale testing programme.

3.4.1. Participants

A total of 56 judges participated. All were first-year undergraduate students completing an 'English Language and Culture' degree at the University of Groningen. They can be considered advanced level students of English since the entry requirements for the course stipulated an IELTS score of 6.5 or more (with a minimum of 6 on all items) or an equivalent. The judging group was composed of several nationalities, but the majority were Dutch nationals. All of the judges were attending a compulsory one-year 'English for Academic Purposes' (EAP) course as part of their main programme and had reached the halfway point at the time of the test. As such, it can be assumed they have some familiarity with the requirements of academic writing in English, though cannot at this stage be considered expert judges.

3.4.2. Design and materials

Each pair was formed from two texts taken from the 12 written by student C. The study adopts a fully-crossed design in which all possible comparisons are made. Each text was, therefore, twice paired with each other text, producing 132 pairs in total. These were then printed onto 66 comparison sheets with one pair (two texts) printed on each side. To avoid participants rating the same text twice in succession, no text which appeared on one side of the sheet was duplicated on the other.

3.4.3. Procedure

The 56 judges were each presented with one comparison sheet (consisting of two pairs) for rating and then 10 of these judges were presented with a further comparison sheet one week later, bringing the total to 66 comparison sheets and 132 pairs rated. As the 132 pairs involved pairs of the same texts in a different order, each pair was rated twice, but by different judges, to increase reliability. A PowerPoint presentation instructed the judges to rate the texts on their use of academic English rather than their content and to ignore the overall structure of the writing, given that they were reading extracts. No further guidance on rating the texts was offered at this point since it was considered important that judges made a holistic judgement of text quality rather than focus on particular lexical or grammatical features. Moreover, since the judges were attending an EAP programme, their notion of 'academic English' should have been sufficiently developed to make meaningful judgements using these criteria.

Judges were advised to spend 2 min reading each text, but were allowed additional time if this was insufficient. They were asked to rate the pairs on both sides of the paper, and to indicate on the sheet which text scored higher on the question: "In terms of the use of academic English, which text do you think is better?" This question was displayed both on the PowerPoint presentation and on the comparison sheet itself. On completion of the test, the number of wins for each text was counted to determine the final rating.

3.5. Data analysis

3.5.1. Research question 1

To answer the first question, the results of each complexity measure taken from the texts of student C will be tested for a Pearson correlation with the text quality scores (the number of 'wins' each text was awarded when compared to the other texts). A significant correlation will be considered an indication of a measure's ability to capture the overall quality of the academic writing.

3.5.2. Research question 2

For the second research question, correlation tests between all six lexical measures and all six syntactic measures will be carried out for each learner. Following Verspoor et al. (2008), detrended data will be used since the general trend of a variable may distort local increases or decreases symptomatic of a competitive relationship with another variable. The data will be detrended using a differencing method which involves creating a new dataset formed of the differences between each successive datapoint in an original series. For example, if the first four measurements for Beyond 2000 are 16.5, 13.7, 17.2 and 26.3 respectively, this will produce a detrended series of −2.8, 3.5 and 9.1, representing the differences between successive datapoints. This results in one fewer datapoint in the final detrended series. A negative correlation between measures shall be considered an indication that the subsystems corresponding to the measures in question are in competition.

4. Results and discussion

4.1. Reliability of the measures

The current study employed six pairs of measures to gauge linguistic complexity, and used a pairwise rating to assess overall text quality. Before presenting the results, this section will inspect the reliability and consistency of these measures.

The internal consistency of each pair of complexity measures designed to tap into the six subconstructs was checked by calculating the correlation coefficient between both members of each pair (see Table 4). This revealed a medium to strong correlation within most pairs indicating that these measures reliably tap into the same construct. A

Table 4
Pearson correlations between complexity measures within constructs.

Construct	Student A	Student B	Student C
Lexical Sophistication	0.71**	0.63**	0.83***
Lexical Diversity	0.93***	0.83***	0.93***
Lexical Density	0.70**	0.76***	0.23
Sentential	0.98***	0.96***	0.99***
Clausal	0.57*	0.83***	0.42
Phrasal	0.52	0.48	0.82**

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

notable exception was a weak correlation ($r(10) = 0.23$, $p = .47$) between Ure and Halliday's measures of lexical density for student C. In this case, the choice of total words or total clauses as the denominator has produced a different outcome. This could suggest, as Halliday (1985) suspected, that the choice of denominator may sometimes be a non-trivial one, although stronger correlations in the case of students A and B indicate the issue is more complex. The sentential measures correlated very strongly for all three students indicating that the use of both word and morpheme variants is probably redundant. This supports findings by Parker and Brorson (2005) who found that the measures tend to be almost perfectly correlated. Similarly, strong correlations between the Guiraud and Vocd-D across the three learners suggest there is little to choose between these two measures, at least between texts of comparable length.

The reliability of the pairwise text quality rating can be considered good since there was a strong correlation ($r(10) = 0.83$, $p < .001$) between the ratings produced each time the pairs were ranked by different judges (each pair was rated twice). This suggests a good degree of consistency across decisions made by the 56 judges.

4.2. Research question 1

To determine which measures of linguistic complexity best capture the overall quality of the academic writing, correlations between the quality ratings and each of the 12 complexity measures were analysed (see Table 5) for the texts of student C. A strong significant relationship (using a Pearson test) was found for five of the measures covering four of the six constructs.

The Beyond 2000 measure correlated most strongly overall, and lexical sophistication proved to be the construct most strongly tied to the text ratings, with both measures highly correlated positively. When measuring lexical density, Halliday's method correlated significantly with text ratings while Ure's method correlated weakly. This is likely a reflection of the weak correlation found between the two measures reported earlier and could look different were the texts from students A and B analysed. Both lexical diversity measures showed a moderate negative correlation with the text ratings, though neither was significant.

Both sentential and phrasal complexity produced medium to strong correlations for both measures, although in each case only one correlation was significant. The correlations for both clausal complexity measures were weak.

The results suggest that in spite of the multi-faceted and dynamic nature of the development of academic writing skills, it is possible to capture text quality with single measures of linguistic complexity. Among these, the Beyond 2000 index best captures the overall quality of academic writing, at least for this learner. This suggests that 'off-list' words, such as those that may be found in the Academic Word List (Coxhead, 2000), have an important role to play on academic writing courses (although Beyond, 2000 does not discriminate between academic words and other off-list words). The findings of Biber and Gray (2010) are also supported by the fact that both phrasal measures correlated well with text ratings while both clausal measures did not. This confirms the view that the use of embedded phrases is an important aspect of academic writing and that this skill deserves attention on academic writing courses. It is also noteworthy that both sentence length measures correlated well. This suggests that despite the structurally compressed nature of academic writing characterised by embedded phrase use, sentential measures are still adequate to capture overall quality. Since the rating of the texts was significantly correlated with the time (number of days since the first text) at which they were written (Pearson's $r(10) = 0.61$, $p < .05$), lexical sophistication and phrasal and sentential complexity might also be the areas of complexity that developed most markedly in this learner's writings.

Table 5
Pearson correlations between complexity measures and text ratings (student C).

Construct	Measure	*r*
Lexical Sophistication	Beyond 2000	0.78**
	P_Lex	0.61*
Lexical Diversity	Guiraud	−0.29
	Vocd-D	−0.36
Lexical Density	Ure's Method	0.23
	Halliday's Method	0.62*
Sentential	ASL (words)	0.57
	ASL (morphemes)	0.62*
Clausal	Clauses per Sentence	−0.09
	Dependent Clauses per Clause	−0.06
Phrasal	Average Clause Length	0.51
	Complex Nominals per Sentence	0.66*

*$p < .05$. **$p < .01$.

4.3. Research question 2

To ascertain whether a competitive relationship exists between lexical and syntactic subsystems, all six lexical measures were compared to all six syntactic measures for each learner and analysed for negative correlations. The results are displayed in Figs. 1–3 which visualise the strength of the correlations using a colour scale. Green is used to indicate a positive correlation, red indicates a negative correlation and, in both cases, the darker the colour, the stronger the correlation.

Since the lexical sophistication, sentential and phrasal measures were found to capture the overall quality of academic writing, it is useful to compare these measures to get a sense of whether there is, overall, a competitive relationship between the lexical and syntactic subsystems of these particular learners. In student B's case, all 4 comparisons between sophistication and sentential measures along with all 4 comparisons between sophistication and phrasal measures were found to correlate negatively (darker red colour in Fig. 2). But in student A's case, a very different pattern emerges with medium to strong positive correlations found between the Beyond 2000 measures and all sentential and phrasal measures (darker green colour in Fig. 1), while little or no correlation can be seen between P_Lex and any of the syntactic measures. In student C's case, there is little evidence of either a positive or negative correlation pattern emerging between any of these constructs (Fig. 3). On this basis, it would seem that in student B's case there is, overall, a competitive relationship between lexical and syntactic subsystems. However, in student A and student C's case, there is no evidence of overall competition with some evidence that there may even be a supportive relationship in student A's case.

Student A			Syntactic Complexity					
			Sentential		Clausal		Phrasal	
			ASL-w	ASL-m	C/S	DC/C	ACL	CN/S
Lexical Complexity	Sophistication	B 2000	0.57	0.63	0.19	0.54	0.45	0.53
		P Lex	0.03	0.10	0.00	0.20	0.02	0.16
	Diversity	Guiraud	-0.18	-0.20	0.24	-0.12	-0.51	-0.43
		Vocd-D	-0.21	-0.25	0.36	-0.19	-0.68	-0.44
	Density	Ure	-0.22	-0.19	-0.44	-0.41	0.24	0.37
		Halliday	0.31	0.32	-0.43	-0.06	0.92	0.56

Note. The colour scale indicates the strength of the correlation.

Fig. 1. Pearson correlations between each of the lexical measures and each of the syntactic measures for student A
Note. The colour scale indicates the strength of the correlation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Student B			Syntactic Complexity					
			Sentential		Clausal		Phrasal	
			ASL-w	ASL-m	C/S	DC/C	ACL	CN/S
Lexical Complexity	Sophistication	B 2000	-0.57	-0.42	-0.09	0.15	-0.47	-0.51
		P Lex	-0.67	-0.61	-0.29	-0.07	-0.36	-0.50
	Diversity	Guiraud	-0.83	-0.82	-0.03	0.12	-0.64	-0.80
		Vocd-D	-0.41	-0.43	-0.10	-0.03	-0.25	-0.55
	Density	Ure	0.45	0.56	-0.31	-0.51	0.61	0.61
		Halliday	0.65	0.72	-0.52	-0.65	0.95	0.75

Note. The colour scale indicates the strength of the correlation.

Fig. 2. Pearson correlations between each of the lexical measures and each of the syntactic measures for student B
Note. The colour scale indicates the strength of the correlation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Student C			Syntactic Complexity					
			Sentential		Clausal		Phrasal	
			ASL-w	ASL-m	C/S	DC/C	ACL	CN/S
Lexical Complexity	Sophistication	B 2000	0.23	0.30	0.01	-0.24	0.21	0.22
		P Lex	0.00	0.08	-0.31	-0.12	0.41	0.08
	Diversity	Guiraud	-0.54	-0.49	0.14	-0.23	-0.81	-0.62
		Vocd-D	-0.45	-0.41	0.27	-0.22	-0.88	-0.59
	Density	Ure	0.11	0.10	0.10	-0.38	-0.10	0.23
		Halliday	0.23	0.19	-0.45	0.03	0.86	0.42

Note. The colour scale indicates the strength of the correlation.

Fig. 3. Pearson correlations between each of the lexical measures and each of the syntactic measures for student C
Note. The colour scale indicates the strength of the correlation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Another pattern that emerges across all three learners is a tendency for the density measures to correlate strongly with phrasal measures, particularly the Halliday and ACL measures. On reflection, a strong correlation here is unsurprising given that Halliday's method counts the number of lexical items per clause (which would normally produce longer clauses) while ACL is a clause length measure. This raises an important question concerning the validity of a lexical density measure. If such a measure always correlates with measures of syntactic complexity and does so by logical necessity, then the extent to which this can be considered a measure of lexical complexity is questionable.

Although the comparison figures do not lend support to the view that there is overall competition between the lexical and syntactic subsystems for all three learners, another interesting feature emerges concerning the relationship between the lexical diversity measures and the sentential and phrasal measures. For students B and C, medium to strong negative correlations exist between the diversity and sentential measures, while a much weaker negative correlation exists for student A. When the diversity and phrasal measures are compared, the negative relationship is even more pronounced with medium to strong negative correlations apparent across all three learners. In Table 6, it can be seen that 7 of the 12 correlations between the diversity and phrasal measures turned out to be significant, lending further support to the view that there is competition between these lexical and syntactic subsystems with respect to these measures.

Importantly, the negative correlations do not point to an increase in one measure over time coupled with a decrease over time in the other measure in any of the three learners. Instead, each measure increases and decreases locally in an alternating fashion with the other measure, resulting in an oscillating pattern (Figs. 4–6). This pattern mirrors that found in Verspoor et al. (2008) when comparing average sentence length and type-token ratio. This suggests that for every text, there is a trade-off and competition for resources between phrasal elaboration and lexical diversification.

What has emerged is that there is some evidence of competition between the lexical and syntactic subsystems in the writing of all three learners over the period of study. In student B's case, there is good reason to believe that, overall, their lexical and syntactic subsystems were in competition since negative correlations were found between the lexical sophistication measures and the sentential and phrasal syntactic measures. Although there was little to suggest that there was, overall, competition between student A and student C's lexical and syntactic subsystems, there was clear evidence of competition between lexical diversity and phrasal measures for all three learners indicating that competition exists between the subsystems associated with these measures.

5. Conclusion

5.1. Summary of findings

The study identified four measures, Beyond 2000, P_Lex, ASL-m and CN/S as strong indices of the overall quality of English academic writing, at least for student C. Halliday's measure of lexical density must be discounted as suitable on the basis that it correlates too strongly with ACL and cannot, therefore, be considered a reliable measure of lexical complexity as distinct from syntactic complexity. The study shows that single measures of complexity can successfully capture academic text quality, but whether the specific measures identified for the one learner here can also do so for texts of other authors must await further corroboration by studies of other individual learners. While this study therefore remains exploratory with respect to which specific aspects of complexity are strongly related to text quality, there is no obvious reason why the quality of texts produced by other learners cannot be rated using the four measures identified.

The second research question concerned whether a competitive relationship existed between the lexical and syntactic subsystems of the three learners. A complex picture emerged in which there was some evidence of overall competition in one of the learners, but not in the other two. However, a competitive relationship was found between the lexical diversity measures and the phrasal syntactic measures from the texts of all three learners. When the correlation between the Guiraud and ACL was traced over time, a repeated competition between these aspects of complexity every time a text was written emerged.

Table 6
Correlations between diversity and phrasal measures.

	Student A		Student B		Student C	
	ACL	CN/S	ACL	CN/S	ACL	CN/S
Guiraud	−0.51	−0.43	−0.64**	−0.80***	−0.81**	−0.62*
Vocd-D	−0.68*	−0.44	−0.25	−0.55*	−0.88***	−0.59

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

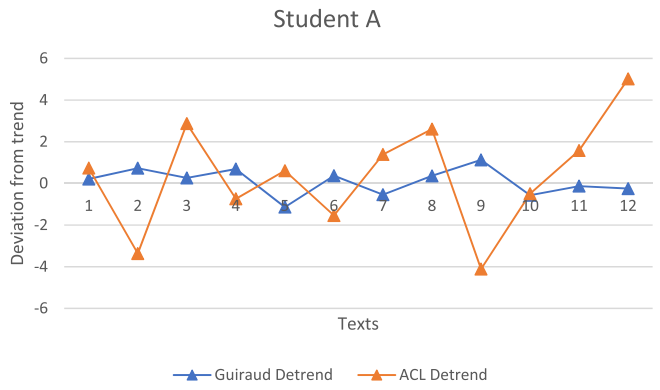

Fig. 4. Detrended representation of Guiraud and Average Clause Length changes over time for student A.

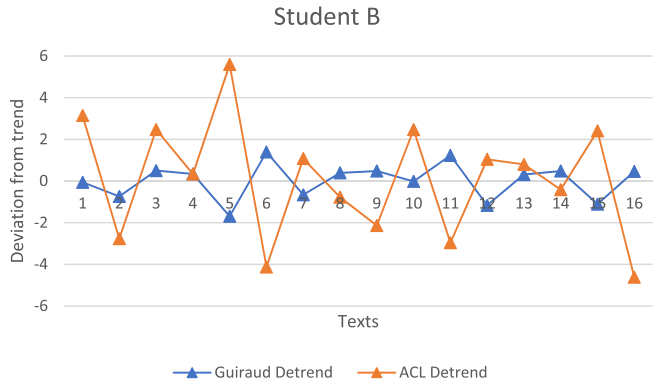

Fig. 5. Detrended representation of Guiraud and Average Clause Length changes over time for student B.

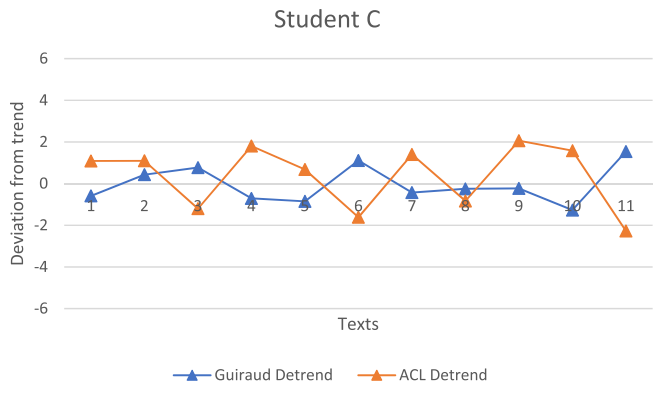

Fig. 6. Detrended representation of Guiraud and Average Clause Length changes over time for student C.

5.2. Limitations and recommendations

The most obvious limitation of any longitudinal study is that it involves the observation of the same set of variables over the period of study without the breadth inherent in a cross-sectional study. Whilst the rationale for using a longitudinal method was made clear at the outset, the limitation remains and further research of a larger number of individuals is needed to confirm the findings. In particular, it seems useful to study the texts of further individuals for correlations between measures of complexity and text quality.

Another limitation of this study is that it did not address the issue of task effect. The essays written by the three participants concerned both literature and linguistics and it was assumed that the tasks were similar enough not to impact the results (Verspoor et al. (2017) found no task effect, for example). It is also possible the difficulty of the essay questions (which presumably increased over time) impacted the kind of language the students produced.

As regards the measures used, the Beyond 2000 measure proved capable of capturing the overall quality of academic writing and so could play a useful role in future research. However, by condensing the LFP in the manner recommended by Laufer (2012), Beyond 2000 is rendered insensitive to academic word usage which may devalue it in relation to academic writing. A better method of condensing the LFP is needed to maximise the utility of the LFP to this type of research.

The lexical density measures did not correlate well together, while Halliday's measure correlated too strongly with ACL suggesting that it taps into a dimension of syntactic rather than lexical complexity. More work is needed to establish whether lexical density should be considered a subconstruct of lexical complexity at all and, if so, what measure should be used to reliably tap into it.

5.3. Practical and theoretical implications

If the four measures identified as reliable indices of academic quality are further validated, this could provide both learners and instructors with a useful tool to estimate the overall quality of a text as well as identify areas for improvement. It may even be possible to combine measures to produce a more reliable overall index. For example, if lexical diversity measures and phrasal measures correlate negatively in most learners at this level, combining these measures could produce a reliable overall index.

The discovery that a competitive relationship existed between the lexical diversity measures and the phrasal syntactic measures was particularly interesting and builds on previous findings by Verspoor et al. (2008) which also found a competitive relationship between these subsystems. Further research is needed to determine whether this is a common pattern amongst other learners. Although it is unlikely from a DST view that this will be the case for every learner, it may be a commonly recurring pattern which could shed light on how resources are shared amongst subsystems at advanced stages of L2 learning.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

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. Lang. Learn. 67, 180–208. https://doi.org/10.1111/lang.12232.

Anthony, L., 2014. AntWordProfiler v1.4.1. Waseda University, Version 1.4.1. http://www.laurenceanthony.net/software.

Biber, D., Gray, B., 2010. Challenging stereotypes about academic writing: complexity, elaboration, explicitness. J. Engl. Acad. Purp. 9, 2–20. https://doi.org/10.1016/j.jeap.2010.01.001.

Biber, D., Gray, B., Poonpon, K., 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? Tesol Q. 45, 5–35. https://doi.org/10.5054/tq.2011.244483.

de Bot, K., Lowie, W., Verspoor, M.H., 2007. A Dynamic Systems Theory approach to second language acquisition. Biling. Lang. Cognit. 10, 7–21. https://doi.org/10.1017/S1366728906002732.

de Bot, K., Larsen-Freeman, D., 2011. Researching second language development from a Dynamic Systems Theory perspective. In: Verspoor, M.H., de Bot, K., Lowie, W. (Eds.),

A Dynamic Approach to Second Language Development: Methods and Techniques. John Benjamins, Amsterdam, pp. 5–24.
Bramley, T., 2007. Paired comparison methods. In: Newton, P., Baird, J.A., Goldstein, H., Patrick, H., Tymms, P. (Eds.), Techniques for Monitoring the Comparability of Examination Standards, Qualifications and Curriculum Authority, pp. 246–300.
Browne, C., Culligan, B., Phillips, J., 2013. The New General Service List. http://www.newgeneralservicelist.org. (Accessed 25 January 2020).
Bulté, B., Housen, A., 2012. Defining and operationalising L2 complexity. In: Housen, A., Kuiken, F., Vedder, I. (Eds.), Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. John Benjamins, Amsterdam, pp. 21–46.
Bulté, B., Housen, A., 2020. A DUB-inspired case study of multidimensional L2 complexity development: competing or connecting growers? In: Lowie, W., Michel, M., Keijzer, M., Steinkrauss, R. (Eds.), Usage-Based Dynamics in Second Language Development. Multilingual Matters, Bristol, pp. 50–86. https://doi.org/10.21832/9781788925259.
Coxhead, A., 2000. A new academic word list. Tesol Q. 34, 213–238. https://doi.org/10.2307/3587951.
Crossley, S.A., McNamara, D.S., 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. J. Sec Lang. Writ. 26, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006.
Douglas, S.R., 2013. The lexical breadth of undergraduate novice level writing competency. Can. J. Appl. Ling. 16, 152–170. https://journals.lib.unb.ca/index.php/CJAL/article/view/21176.
Flesch, R., 1948. A new readability yardstick. J. Appl. Psychol. 32, 221–233. https://doi.org/10.1037/h0057532.
Gebril, A., Plakans, L., 2016. Source-based tasks in academic writing assessment: lexical diversity, textual borrowing and proficiency. J. Engl. Acad. Purp. 24, 78–88. https://doi.org/10.1016/j.jeap.2016.10.001.
van Geert, P., 2008. The dynamic systems approach in the study of L1 and L2 acquisition: an introduction. Mod. Lang. J. 92, 179–199. https://doi.org/10.1111/j.1540-4781.2008.00713.x.
Guiraud, P., 1954. Les caracteres statistiques du vocabulaire: Essai de methodologie. Presses Universitaires de France, Paris.
Halliday, M.A.K., 1985. Spoken and Written Language. Deakin University Press, Geelong.
Heldsinger, S., Humphry, S., 2010. Using the method of pairwise comparison to obtain reliable teacher assessments. Aust. Educ. Res. 37, 1–19. https://doi.org/10.1007/BF03216919.
Higginbotham, G., Reid, J., 2019. The lexical sophistication of second language learners' academic essays. J. Engl. Acad. Purp. 37, 127–140. https://doi.org/10.1016/j.jeap.2018.12.002.
Huddleston, R., Pullum, G.K., 2005. A Student's Introduction to English Grammar. Cambridge University Press, London.
Hunt, K.W., 1965. Grammatical Structures Written at Three Grade Levels. National Council of Teachers of English, Champaign, IL.
Ishikawa, T., 2007. The effect of manipulating task complexity along the [+/- Here-and-Now] dimension on L2 written narrative discourse. In: García Mayo, M. (Ed.), Investigating Tasks in Formal Language Learning, Multilingual Matters, Clevedon, pp. 136–156.
Klee, T., Fitzgerald, M.D., 1985. The relation between grammatical development and mean length of utterance in morphemes. J. Child Lang. 12, 251–269. https://doi.org/10.1017/S0305000900006437.
Kyle, K., 2016. Measuring syntactic development in L2 writing: fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Ph.D. dissertation). http://scholarworks.gsu.edu/alesl_diss/35.
Larsen-Freeman, D., 1997. Chaos/complexity science and second language acquisition. Appl. Ling. 18, 141–165. https://doi.org/10.1093/applin/18.2.141.
Laufer, B., 2012. Lexical frequency profiles, the encyclopedia of applied linguistics. https://doi.org/10.1002/9781405198431.wbeal0692 (accessed 25 January 2020).
Laufer, B., Nation, P., 1995. Vocabulary size and use: lexical richness in L2 written production. Appl. Ling. 16, 307–322. https://doi.org/10.1093/applin/16.3.307.
Laufer, B., 1995. Beyond 2000. In: Eubank, L., Selinker, L., Smith, M.S., Rutherford, W.E. (Eds.), The Current State of Interlanguage. John Benjamins, Amsterdam, pp. 265–272.
Lemmouh, Z., 2008. The relationship between grades and the lexical richness of student essays. Nordic J. Engl. Stud. 7, 163–180. https://doi.org/10.35360/njes.106.
Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. Int. J. Corpus Linguist. 15, 474–496. https://doi.org/10.1075/ijcl.15.4.02lu.
MacWhinney, B., Snow, C., 1990. The child language data exchange system: an update. J. Child Lang. 17, 457–472. https://doi.org/10.1017/S0305000900013866.
Malvern, D.D., Richards, B.J., 1997. A new measure of lexical diversity. In: Ryan, A., Wray, A. (Eds.), Evolving Models of Language, Multilingual Matters, Clevedon, pp. 58–71.
McKee, G., Malvern, D.D., Richards, B.J., 2000. Measuring vocabulary diversity using dedicated software. Lit. Ling. Comput. 15, 323–338. https://doi.org/10.1093/llc/15.3.323.
Meara, P., 2018. P_Lex v3.00. _lognostics, Version 3.00. http://www.lognostics.co.uk/tools/index.htm.
Meara, P., Bell, H., 2001. P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. Prospect 16, 5–17.
Morris, L., Cobb, T., 2004. Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. System 32, 75–87. https://doi.org/10.1016/j.system.2003.05.001.
Nation, P. (Ed.), 1984. Vocabulary Lists: Words, Affixes, and Stems. ELI Occasional Publications, Wellington.
Norris, J.M., Ortega, L., 2009. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. Appl. Ling. 30, 555–578. https://doi.org/10.1093/applin/amp044.
[computer code] O'Leary, J.A., 2022. Python Program for Extracting Representative Text Samples. Github. https://github.com/johnol22/Python-program-for-extracting-representative-text-samples.
O'Loughlin, K., 1995. Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. Lang. Test. 12, 217–237. https://doi.org/10.1177/026553229501200205.
Pallotti, G., 2015. A simple view of linguistic complexity. Sec. Lang. Res. 31, 117–134. https://doi.org/10.1177/0267658314536435.
Parker, M.D., Brorson, K., 2005. A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). First Lang. 25, 365–376. https://doi.org/10.1177/0142723705059114.
Parkinson, J., Musgrave, J., 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. J. Engl. Acad. Purp. 14, 48–59. https://doi.org/10.1016/j.jeap.2013.12.001.
Ravid, D., Zilberbuch, S., 2003. The development of complex nominals in expert and non-expert writing. Pragmat. Cognit. 11, 267–296. https://doi.org/10.1075/pc.11.2.05rav.
Sercu, L., de Wachter, L., Peters, E., Kuiken, F., Vedder, I., 2006. The effect of task complexity and task conditions on foreign language development and performance: three empirical studies. ITL - Int. J. Appl. Linguist. 152, 55–84. https://doi.org/10.2143/ITL.152.0.2017863.
Skehan, P., 2009. Modelling second language performance: integrating complexity, accuracy, fluency, and lexis. Appl. Ling. 30, 510–532. https://doi.org/10.1093/applin/amp047.
Storch, N., Wigglesworth, G., 2007. Writing tasks: the effects of collaboration. In: García Mayo, M. (Ed.), Investigating Tasks in Formal Language Learning, Multilingual Matters, Clevedon, pp. 157–177.
Templin, M.C., 1957. Certain Language Skills in Children: Their Development and Interrelationships. University of Minnesota Press, Minneapolis.
Thurstone, L.L., 1927. A law of comparative judgment. Psychol. Rev. 34, 273–286. https://doi.org/10.1037/h0070288.
To, V., Fan, S., Thomas, D., 2013. Lexical density and readability: a case study of English textbooks. Internet J. Lang. Cult.Soc. 37, 61–71. https://aaref.com.au/wp-content/uploads/2018/05/37-07.pdf.
Ure, J.N., 1971. Lexical density and register differentiation. In: Perrin, G.E., Trim, J.L.M. (Eds.), Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics. Cambridge University Press, London, pp. 443–452.
Verspoor, M.H., Lowie, W., van Dijk, M., 2008. Variability in second language development from a dynamic systems perspective. Mod. Lang. J. 92, 214–231. https://doi.org/10.1111/j.1540-4781.2008.00715.x.
Verspoor, M.H., Lowie, W., Chan, H.P., Vahtrick, L., 2017. Linguistic complexity in second language development: variability and variation at advanced stages. Recherches en Didactique des Langues et des Cultures 14, 1–27. https://doi.org/10.4000/rdlc.1450.
West, M., 1953. A General Service List of English Words. Longman, London.
Yixin, W., Daller, M.H., 2014. Predicting Chinese students' academic achievement in the UK. In: Proceedings of the 47th Annual Meeting of the British Association for Applied Linguistics: Learning, Working and Communicating in a Global Context. Scitsiugnil Press, London.
Zenker, F., Kyle, K., 2021. Investigating minimum text lengths for lexical diversity indices. Assess. Writ. 47, 100505 https://doi.org/10.1016/j.asw.2020.100505.