System xxx (xxxx) 102246

Contents lists available at ScienceDirect

System

journal homepage: http://ees.elsevier.com

Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech

Bram Bulté [a,*], Hanne Roothooft [b]

[a] CLIN, Vrije Universiteit Brussel, Pleinlaan 2, 1040, Brussel, Belgium
[b] Universidad Pública de Navarra, Spain

ARTICLE INFO

Article history:
Received 3 September 2019
Received in revised form 23 March 2020
Accepted 25 March 2020
Available online xxx

Keywords:
Syntactic complexity
Lexical complexity
Morphological complexity
Proficiency
Speech
IELTS

ABSTRACT

This study investigates the relationship between nine quantitative measures of L2 speech complexity and subjectively rated L2 proficiency by comparing the oral productions of English L2 learners at five IELTS proficiency levels. We carry out ANOVAs with pairwise comparisons to identify differences between proficiency levels, as well as ordinal logistic regression modelling, allowing us to combine multiple complexity dimensions in a single analysis. The results show that for eight out of nine measures, targeting syntactic, lexical and morphological complexity, a significant overall effect of proficiency level was found, with measures of lexical diversity (i.e. Guiraud's index and HD-D), overall syntactic complexity (mean length of AS-unit), phrasal elaboration (mean length of noun phrase) and morphological richness (morphological complexity index) showing the strongest association with proficiency level. Three complexity measures emerged as significant predictors in our logistic regression model, each targeting different linguistic dimensions: Guiraud's index, the subordination ratio and the morphological complexity index.

1. Introduction

One of the main questions in second language acquisition (SLA) research relates to the nature of the (linguistic) changes taking place in the second language (L2) system of the learners as they develop or, in other words, as they become more proficient. L2 production samples can be analysed as concrete manifestations of what learners are capable of doing at a precise point in their development, providing a snapshot of the current state of the more abstract, underlying L2 system, which cannot be studied directly. The type of empirical evidence thus gathered is not only relevant from a descriptive point of view, but can also contribute to the formulation of explanatory hypotheses regarding second language development. Moreover, being able to objectively measure progress, development or proficiency in a second language is of (practical) interest not only for SLA researchers, but also for language testers and educational professionals.

The task of assessing L2 proficiency or L2 development through an objective analysis of L2 production data has been approached in different ways. Researchers have for example attempted to identify linguistic features that are characteristic of L2 production at different L2 proficiency levels. A major endeavour in this respect is the English Profile project, of which one of the aims was to match different levels of the Common European Framework of Reference for languages (CEFR) with English lexical and grammatical features (Hawkins & Filipović, 2012). A similarly data-driven approach is taken by studies that apply techniques developed in the context of Biber's (1988) Multidimensional Analysis to second language development data (e.g. Crosthwaite, 2016; Friginal & Weigle, 2014). The idea here is to verify which linguistic features cluster together and appear more frequently in the productions of more (or less) advanced learners.

Our study fits within the complexity-accuracy-fluency approach to assessing L2 development and proficiency (Ellis & Barkhuizen, 2005; Housen, Kuiken & Vedder, 2012; Michel, 2017). These three constructs are claimed to represent crucial axes along which lan-

* Corresponding author.
E-mail addresses: bram.bulte@vub.be (B. Bulté); hanne.roothooft@unavarra.es (H. Roothooft)

https://doi.org/10.1016/j.system.2020.102246
0346-251/

guage learning can be traced by studying L2 productions. While fluency and accuracy deal with phenomena such as speed and correctness;,complexity, generally speaking, focuses on aspects such as the range and diversity of linguistic forms that appear in a learner's production (Bulté & Housen, 2012; Ortega, 2003). In the present study, we focus on L2 complexity. As we will discuss below, complexity has been interpreted in a number of different ways in the SLA literature. We adhere to a narrow, formal definition of complexity, which does not make reference to subjective notions such as difficulty or empirical observations of L2 development.

A considerable number of studies have analysed the relationship between L2 writing complexity and proficiency (see e.g. Ortega, 2003 and Wolfe-Quintero, Inagaki, & Kim, 1998 for research syntheses, and Abbas Khushik & Huhta, 2019; Gyllstad, Granfeldt, Bernardini, & Källkvist, 2014; Lu, 2011; Verspoor, Schmid, & Xu, 2012; Yang, Lu, & Weigle, 2015; Yoon, 2017 for recent studies). However, we cannot assume that findings from writing research can be generalised to oral proficiency, as there tend to be important differences between the written and oral mode (Biber, Gray, & Staples, 2016). For instance, in a recent study by Vasylets, Gilabert, and Manchón (2019), learners who performed the same video-retelling task in the written and oral mode used syntactically and lexically more complex language in writing. Compared to the work on written language, there are relatively few studies which have analysed complexity at different levels of oral proficiency (see literature review). Moreover, these studies tend to focus on one linguistic domain only (i.e. either syntax or vocabulary), with only a few of them targeting morphological complexity. Complexity measures are typically considered in isolation, and the studies often do not provide detailed information about which measures correlate with or distinguish between proficiency levels. To address these gaps in previous research, the aim of this study is to further investigate the empirical connection between formal complexity and L2 proficiency by examining whether quantitative measures of syntactic, lexical and morphological complexity applied to L2 oral production data can discriminate between learners that have been placed at different levels of the IELTS English speaking test. We also test which (combination of) complexity measures best correlates with proficiency. Our study is cross-sectional in nature, and therefore does not investigate L2 development as such, unlike a number of recent studies that have been conducted in a complex/dynamic systems theory framework (Bulté & Housen, 2018; Spoelman & Verspoor, 2010; Vyatkina, Hirschmann, & Golcher, 2015). Our aim, rather, is to capture L2 development in broad strokes by looking at similarities and differences in the L2 productions of learners at different, broadly defined proficiency levels (see Verspoor et al., 2012).

2. Literature review

2.1. L2 complexity

The term complexity has taken on a number of closely related meanings in the context of SLA research, which has led to some terminological confusion (Bulté & Housen, 2012, 2014; Ortega, 2012; Pallotti, 2009, 2015). For example, Ellis and Barkhuizen (2005) define complexity as the "use of more challenging and difficult language" (Ellis & Barkhuizen, 2005, p. 139), while Biber, Gray, and Poonpon (2011) refer to "the more advanced grammatical structures that students exhibit as they progress in their language proficiencies" (p. 6). Whereas the first definition equates complexity with difficulty, a subjective notion related to cognitive processing, the second defines complexity in terms of empirical observations of the L2 production of learners with different levels of proficiency. In this article, we adhere to a formal, objective and ultimately quantitative definition of complexity. We define complexity in terms of the number and diversity of elements that a structure or a system comprises, as well as the number of (hierarchical) relationships between these elements (Bulté & Housen, 2012; Pallotti, 2015; Rescher, 1998). Defined in this way, complexity manifests itself in L2 production in, for example, the number, range and diversity of lexical items, syntactic structures and grammatical morphemes, the length or compositionality of these items and structures, and the degree of embedding or the number of dependency relationships between them.

2.2. Operationalising L2 proficiency

The notion of L2 proficiency has also been interpreted in different ways in L2 research. Broadly speaking, proficiency can be defined as the overall level of development of an L2 learner, meaning how well the learner knows a language, or how well she is able to use the language in diverse communicative situations (Leclercq, Edmonds & Hilton, 2014; see Hulstijn, 2011, for a thorough discussion of how to define L2 proficiency). Usually proficiency is defined rather loosely or even implicitly, and its operationalisation differs greatly across L2 studies (Ortega, 2003; Thomas, 1994). This is also the case for studies that have investigated the linguistic features that correlate with or distinguish between L2 proficiency levels. A number of studies have taken students' course level as a measure of their proficiency (e.g., Lu, 2011), while others use holistic rating scales, either designed by researchers (Connor-Linton & Polio, 2014), or used in standardised language tests such as TOEFL (Iwashita, Brown, McNamara, & O'Hagan, 2008) or IELTS (Read & Nation, 2006; Seedhouse, Harris, Naeb, & Üstünel, 2014). A number of studies have also operationalised proficiency in terms of CEFR levels (Abbas Khushik & Huhta, 2019; Gyllstad et al., 2014). In a meta-analysis of L2 writing studies, Ortega (2003) observed that proficiency levels based on holistic rating yielded more homogeneous samples than those based on programme level.

2.3. L2 speech complexity and L2 proficiency

A considerable number of studies investigated the relationship between L2 proficiency and L2 writing complexity, but for the purpose of our study, we are mainly interested in studies targeting L2 speech. Even though some studies have investigated complexity

in languages other than English, such as L2 Japanese (Iwashita, 2006), L2 German (Neary-Sundquist, 2017) and L2 French (De Clercq, 2015; De Clercq & Housen, 2017, 2019), the focus in this study is on L2 English, as has been the case for most studies so far. Table 1 provides an overview of nine studies that have investigated the relationship between complexity and oral proficiency for English as a second or foreign language, indicating the number of participants per study, as well as the number and range of proficiency levels and the way in which proficiency and complexity are operationalised.[1] As can be seen in Table 1, the first three studies focused exclusively on lexical complexity, three studies investigated syntactic complexity, two studies included measures of both lexical and syntactic complexity and only one study looked at morphological complexity.

The first article addressing the relationship between oral proficiency and lexical complexity reports on a comparison between English and French as a foreign language in spoken data gathered from 100 secondary school learners with L1 Dutch (De Clercq, 2015). The texts were divided into four proficiency levels based on the learners' age and their level of linguistic accuracy in the L2. De Clercq (2015) studied the development of lexical complexity in these four levels and he concluded that the measures of lexical diversity (D, Guiraud's index) were best able to distinguish between proficiency levels. Both D and Guiraud's index were able to differentiate between adjacent levels 1 and 2, and 2 and 3, but no differences were found between levels 3 and 4 (i.e. the two highest levels).

Lu (2012) analysed three types of lexical complexity, using a wide range of measures for each, in 408 oral narratives performed by Chinese college-level learners of English, who were rated on a holistic rubric and grouped into four proficiency levels. The author used computer-based analyses to calculate measures of lexical density, lexical sophistication and lexical variation. Only lexical variation appeared to be a good predictor of proficiency, which is in line with De Clercq's (2015) findings. Even though Lu (2012) performed ANOVAs to investigate whether measures yielded significant differences between levels, the author did not carry out post-hoc analyses to find out where the differences were located.

Of the two studies that looked at complexity and oral proficiency using data from the IELTS speaking test, one focused on lexical complexity. Read and Nation (2006) calculated the overall number of tokens and types, lexical variation (D) and lexical sophistication (lambda) for 88 samples ranging from band 4 to band 8. In general, the number of words produced, lexical variation, and the number of infrequent or sophisticated words increased gradually from bands 4 to 8, but there was a lot of variation within each level and the authors did not perform statistical tests to ascertain if these differences were significant, or how well the chosen measures could distinguish between levels.

With regard to syntactic complexity, De Clercq and Housen (2017) analysed the same dataset as De Clercq (2015) and found that length of AS-unit and clauses per AS-unit were the measures that most clearly distinguished between proficiency levels, while two of the diversity measures also yielded significant differences between levels (i.e. AS-unit diversity and Syntactic Diversity Index).

Gan (2012) also analysed five measures of syntactic complexity in two different oral exam tasks that were part of a national English test in Hong Kong. Students had to perform a monologue and a group discussion, which were rated by their teacher based on analytical scales, resulting in six proficiency levels. The only measure that correlated significantly with the global test scores was mean length of utterance.

The only study, to the best of our knowledge, which has investigated syntactic complexity at different levels of the IELTS speaking test, is Seedhouse et al. (2014). These authors only used two measures of syntactic complexity, both targeting clause-linking by means of subordination: A-units per AS-unit and A-units per total number of words. Foster, Tonkyn, and Wigglesworth (2000, p.366) define A-units as "subordinate clauses which have at least a finite or non-finite verb element plus at least one other clause element such as subject, object, complement, or adverbial". Both measures gave rise to statistically significant differences between bands, but no information is included about where the differences are located.

Iwashita et al. (2008) studied a number of measures of accuracy, fluency and complexity in five different levels of the TOEFL iBT speaking test. The complexity metrics were mainly syntactic measures, even though one measure of lexical complexity was also included. For syntactic complexity, significant differences between levels were only observed for verb phrases per T-unit and mean length of utterance. With regard to lexical complexity, the number of types and tokens significantly increased from level 1 to level 5. However, no information is provided about where exactly the differences lie, and whether or not these measures can discriminate between adjacent levels.

Kang (2013) investigated which features could distinguish between different levels of the Cambridge English speaking tests (B1 to C2), focusing on fluency, pronunciation, and grammatical and lexical complexity. For lexical complexity, significant differences between levels were found for the type-token ratio, even though the increase was not linear from the lowest to the highest level. For the lexical measures, differences were mainly located between B1 and B2, and between C1 and C2, but not between B2 and C1. All measures of syntactic complexity under investigation, except the number of T-units, gave rise to significant differences, with syntactic complexity increasing from the B1 to the C2 level. Moreover, total number of clauses, total number of dependent clauses and clauses per T-unit were found to distinguish between B1, B2 and C2.

Finally, we only found one study that investigated the relationship between morphological complexity and oral proficiency, De Clercq and Housen (2019), who analysed three measures of morphological complexity in the same oral narratives that were used by De Clercq (2015) and De Clercq and Housen (2017). The measures analysed were types per family (Horst & Collins, 2006),

[1] The studies by De Clercq (2015) and De Clercq and Housen (2017, 2019) compare complexity in English and French monologues performed by the same learners, but for the present study we mainly focus on the results for English.

Table 1
Key characteristics of nine studies investigating the relationship between L2 speech complexity and L2 proficiency.

Study	Learners	Number of proficiency levels	Range of proficiency levels	Proficiency assessment	Complexity measures
De Clercq (2015)	100	4	Beginner - advanced	Age and accuracy	- Lexical variation (D) - Guiraud's index for nouns and verbs - Word frequency and reaction time - Lexemic density - Collocational density
Lu (2012)	408	4	General labels only ("fail", "pass", "good", "excellent")	Holistic rubric	- Lexical density - 5 lexical sophistication measures - 20 lexical variation measures
Read and Nation (2006)	88	5	Band 4 to band 8	IELTS	- # tokens - # types - Lexical variation (D) - Lexical sophistication (lambda)
De Clercq and Housen (2017)	100	4	Beginner - advanced	Age and accuracy	- Mean length of noun phrase, clause, AS-unit - Clauses per AS-unit - Syntactic diversity
Gan (2012)	30	6	No information	Hong Kong national English test scores	- Mean length of T-unit - Clauses per T-unit - Dependent clauses per clause - Verb phrases per T-unit - Mean length of utterance
Seedhouse et al. (2014)	60	4	Band 5 to band 8	IELTS	- A-units per AS-unit - A-units per total number of words
Iwashita et al. (2008)	200	5	No information	TOEFL iBT scores	- Clauses per T-unit - Dependent clauses per clause - Verb phrases per T-unit - Mean length of utterance - # tokens - # types

Table 1 (*Continued*)

Study	Learners	Number of proficiency levels	Range of proficiency levels	Proficiency assessment	Complexity measures
Kang (2013)	120	4	B1 to C2	Cambridge English	- Type-token ratio - Lexical density - Lexical sophistication - # T-units, clauses, dependent clauses - Clauses per T-unit
De Clercq and Housen (2019)	100	4	Beginner - advanced	Age and accuracy	- Types per family - Inflectional diversity - Morphological complexity index

Malvern, Chipere, Richards and Durán's (2004) Inflectional Diversity and Pallotti's (2015) Morphological Complexity Index. For the English data, these measures could only discriminate between the lowest level, on the one hand, and the three highest levels, on the other.

In conclusion, the studies of the relationship between L2 proficiency and L2 speaking complexity reviewed here provide a rich and varied picture that is not easily summarised due to differences in research design relating to the operationalisation of both complexity and proficiency, the sample of L2 learners analysed and the statistical analyses. In spite of these methodological differences, a number of trends can be distilled from the studies discussed above. With regard to lexical complexity, measures of lexical diversity (Guiraud's index and D in particular) appear to be best able to distinguish between proficiency levels, and potentially also between adjacent levels. At the level of syntactic complexity, especially those measures targeting the average length of supra-clausal units (i.e. T-units, AS-units or utterances) and those tapping into clausal subordination (e.g. clauses per AS-unit or T-unit) have been shown to increase with increasing proficiency. The only evidence regarding morphological complexity seems to suggest that measures targeting this aspect of complexity are only able to distinguish between learners at low proficiency levels (in L2 English).

3. Research questions

Even though a number of studies have investigated the relationship between complexity and oral proficiency in L2 English, we have seen that most of these studies tend to focus either on lexical complexity or syntactic complexity, and only one study has investigated morphological complexity. Few studies have also provided information about which measures distinguish between which proficiency levels, or which combination of measures best predicts proficiency. The present study seeks to address this gap by analysing a range of syntactic, lexical and morphological complexity measures at five different proficiency levels, investigating which measures best discriminate between proficiency levels and which combination of syntactic, lexical and morphological measures can be used to predict proficiency level.

Thus, the present study addresses the following broad research questions:

1. Can measures of lexical, syntactic and morphological complexity applied to oral L2 English productions discriminate between learners at different L2 proficiency levels?
   a. Which measures are best able to discriminate between proficiency levels?
   b. Between which specific proficiency levels are the differences situated?
2. Do measures of lexical, syntactic and morphological complexity correlate with L2 proficiency?
   a. Which individual complexity measures correlate best with L2 proficiency?
   b. Which combination of complexity measures best predicts L2 proficiency level?

4. Material and methods

4.1. Dataset and participants

Our dataset consists of 71 samples from the *International English Language Testing System (IELTS*) speaking test, pertaining to five adjacent proficiency levels. The IELTS exam is a high-stakes English test developed by the British Council and Cambridge Assessment English. Learners who take the IELTS exam do so mainly for immigration and study purposes (certain universities in e.g. the UK and Australia require a minimum IELTS score to enter). The IELTS exam measures English language proficiency on a scale that ranges from band 1 to band 9.[2] The learners in this study belong to band 4 (14 samples), band 5 (14 samples), band 6 (15 samples), band 7 (14 samples) and band 8 (14 samples). These band levels have been (loosely) related to the proficiency levels described in the Common European Framework of Languages.[3] The samples in our corpus are produced by learners pertaining to CEFR levels A2/B1 to C1. The corpus consists of one sample per participant, and the participants come from different first-language backgrounds, with some languages being more frequent than others, such as Arabic, Chinese or Tagalog.

The speaking test takes the form of an interview between a candidate and an examiner, which lasts for 11–14 minutes. In the second part of the test, the candidate is required to speak about a topic for up to two minutes. It is this monologue or long turn which we analysed for the present study. Candidates are given a topic and they receive one minute to prepare what they are going to say. Examples of topics are: "describe one of your neighbours", "describe some travelling you would like to do in the future" or "describe a building that you like". The speaking test is scored by means of a rubric and the band level is calculated by averaging the scores on four different aspects: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation.

4.2. Complexity measures

We calculated a total of nine complexity measures tapping into syntactic (5), lexical (3), and morphological complexity (1). Table 2 provides an overview of the measures and their calculation.

The syntactic complexity measures target the following, hierarchically structured, production units: AS-units, clauses and phrases. Clauses are combined into AS-units by means of subordination; in turn, noun phrases are a constituent component of clauses. All syntactic measures calculated in this study are operationalised as ratios, and target either the length of production units (i.e. number of words per AS-unit, clause or noun phrase) or the proportion of certain specific structures relative to their 'parent' category (i.e. number of subclauses or coordinated clauses divided by the total number of clauses).

Mean length of AS-unit (MLAS) is the most global syntactic measure included in this study, in that AS-units can be made longer by adding more clauses together, or by lengthening the constituent clauses or phrases. An AS-unit is defined as "a single speaker's utterance consisting of an *independent clause, or sub-clausal unit*, together with any *subordinate clause(s)* associated with either" (Foster et al., 2000, p. 365). AS-units are arguably more adapted for analysing spoken language than T-units, which are often used in the analysis of written data. Also for the identification of clauses and subclauses we follow the guidelines provided by Foster et al. (2000). For the coordination ratio, we only count independent coordinated clauses. More details about the calculation of mean noun phrase length are provided in Bulté & Housen, 2014.

[2] In practice, IELTS does not use levels 1–3.

[3] https://www.ielts.org/ielts-for-organisations/common-european-framework.

Table 2
Complexity measures.

Measure	Calculation
Syntactic complexity	
Mean length of AS-unit (MLAS)	words/AS-units
Subclause ratio (SCR)	subclauses/clauses * 100
Mean length of clause (MLC)	words/clauses
Coordination ratio (CCR)	independent coordinated clauses/clauses
Mean length of noun phrase (MLNP)	words in noun phrases/noun phrases
Lexical complexity	
Guiraud's index (G)	types/V tokens
Hypergeometric distribution (HD-D)	see McCarthy and Jarvis (2007)
Measure of textual lexical diversity (MTLD)	see McCarthy and Jarvis (2010)
Morphological complexity	
Morphological complexity index - verbs (MCI)	see Brezina and Pallotti (2019)

The three measures of lexical complexity are all variations of the lexical type-token ratio (TTR) that try to compensate for unwanted text-length effects (i.e. the longer a text is, the higher the probability that words are repeated). Guiraud's index offers the most simple transformation, in that a square root is added to the denominator of the simple TTR formula (Guiraud, 1959). This transformation actually overcompensates for the effects of text length, and longer texts tend to get higher scores on Guiraud's index than shorter texts (Bulté, Housen, Pierrard, & Van Daele, 2008). Thus, Guiraud's index taps into three distinct aspects of lexical diversity, namely size (i.e. word tokens), richness (i.e. word types) and the effective number of types (Jarvis, 2013). MTLD and HD-D both only target "the range of different words in a text" (McCarthy & Jarvis, 2010, p. 381) or, in other words, the effective number of types, by relying on a more complicated, and mathematically speaking more correct transformation of the TTR to compensate for text length effects (McCarthy & Jarvis, 2007, 2010).

Morphological complexity is measured by means of the morphological complexity index (MCI) applied to verbs (Brezina & Pallotti, 2019). This index, whose underlying logic is similar to that of lexical TTRs, targets the inflectional diversity of verb forms in a text by analysing the number of different morphological processes (or 'exponents') that are applied. Its computation is based on repeated random sampling from all exponents in a text, and counting the average number of different exponents within and across these samples.[4] By keeping the size of the random samples constant, unwanted text length effects are eliminated. An MCI value of 0 means that all verb forms in the text are the same, morphologically speaking, and higher values indicate that a text contains more different morphological verb forms. The theoretical maximum value for this measure, which depends on the selected sample size, is 9 in this study.

4.3. Data coding and analysis

Nearly all audio files provided by IELTS were already accompanied by transcriptions, and the remaining ones were transcribed by the second author. The transcriptions were quite detailed, as they contained information about, for example, sentence stress, pause length, hesitations and false starts. Data preparation consisted in removing hesitation markers ("erm") and interjections such as "yeah, you know, well", false starts (for instance: "as Asian we are free to go, to enter Hong Kong" becomes: "as Asian we are free to enter Hong Kong") and repetitions (for instance "It's a lot of fun when you, when you travel"). Some inaudible or unclear fragments also had to be left out.

In a next step, the samples were divided into AS-units (see definition above) and clauses. Following Foster et al. (2000), we considered non-finite clauses to be separate clauses if they contained a verb and at least one other element. For example: "It's a good place to live (band 5)" was analysed as one clause, while "One day she ask me/to help her (band 6)" was divided into two clauses. The texts were manually coded for syntactic features by the second author, in case of doubt after discussion with the first author. To establish inter-coder reliability, 15 randomly selected texts (i.e. slightly over 20% of the corpus) were coded separately by the first author. The correlations between the two codings ranged from 0.852 for MLNP to 0.983 for MLC.[5] The tool for the automatic analysis of lexical diversity (TAALED[6]) was used for the lexical analyses, and the MCI was calculated with the computer tool developed by Brezina and Pallotti (2019).[7]

[4] We work with 100 random samples of 5 exponents each.
[5] SCR: 0.931; MLAS: 0.950; CCR: 0.976.
[6] www.linguisticanalysistools.org, version 1.2.4.
[7] http://corpora.lancs.ac.uk/vocab/analyse_morph.php, alpha version.

4.4. Statistical analyses

We used ANOVAs with 10 pairwise comparisons (least significant difference) to assess differences between the five IELTS proficiency levels. Levene's test showed that only for one of the measures (mean length of AS-unit) the assumption of homogeneity of variances was not met ($p = 0.014$), mainly due to a much larger variability in the highest proficiency level.[8] For the sake of consistency, we decided to use standard ANOVAs for each of the measures, also considering the very similar group sizes, which positively affects the overall robustness of the analyses. The main ANOVAs were used to estimate whether the overall effect of proficiency level on the various complexity scores was significant and what the size of this effect was (measured with eta^2). The pairwise comparisons provide more information about the specific location of the effect. We do not apply corrections for multiple testing (e.g. Bonferroni) since the comparisons are planned and not meant for hypothesis testing in a narrow sense. In addition, the size of our subsamples per proficiency level is relatively small, yet almost identical. As a result, the reported p-values should not be interpreted absolutely, but rather as providing indications of the precise location of sizeable differences between proficiency levels.[9]

In a second step, we use correlation analyses as well as ordinal logistic regression to ascertain which (combinations of) measures show the strongest relationship with proficiency level. We use Spearman rank correlations to test the strength of the association between proficiency level and each of the complexity measures. We then fit a cumulative logit model with proportional odds to the data using PROC LOGISTIC in SAS. Such a model imposes equal slopes to be estimated for each response function; only the intercepts are allowed to vary, even though they are constrained to increase (Agresti, 2010; Derr, 2013). In other words, the effects of the predictors are assumed to be the same for each adjacent pair of proficiency levels. The proportional odds assumption is tested by means of a score test. We report Nagelkerke's R^2 as an approximation of the R^2 statistic used to indicate the proportion of explained variance in linear regression models. Proficiency level is used as dependent variable in this ordinal logistic regression model, and complexity measures are entered as independent variables. To avoid multicollinearity, highly correlating complexity measures (i.e. $r > 0.8$) are left out of the model.

5. Results

5.1. Can complexity measures discriminate between L2 proficiency levels?

In this section we report the results of the ANOVAs and pairwise comparisons. The descriptive statistics for all measures (mean and standard deviation per proficiency level and overall) are provided in Appendix A; Appendix B contains an overview of the results of the ANOVAs. For each complexity measure, we summarise the results by means of a bar chart showing the mean score per proficiency level (surrounded by a 95% confidence interval), as well as potential significant differences between pairs of proficiency levels (10 pairs in total per analysis), indicated in the top part of the graph.

Overall, mean AS-unit length differed significantly across proficiency levels, $F(4,66) = 8.574$, $p < .001$, $\eta^2 = 0.342$. Fig. 1 shows that the mean length of AS-units steadily increases from around nine and a half words in level 4 to over 15 words in level 8. Seven out of 10 pairwise comparisons showed sizeable differences between proficiency levels. Yet, only for one pair of adjacent proficiency levels (levels 6 and 7) a sizeable difference was found.

Also for the subclause ratio the scores differed significantly across proficiency levels, $F(4,66) = 3.714$, $p = .009$, $\eta^2 = 0.184$. Fig. 2 shows that around 39% of the clauses produced by the learners at level 4 was subordinated. By level 8, this proportion had increased to over 56%. Even though there is a clear increasing trend in SCR scores with increasing proficiency level, only four pairwise differences between levels were found to be significant. No sizeable differences between adjacent levels were observed. The scores for the learners in the highest proficiency level differed considerably from the scores in all levels, except for level 7.

The mean clause length differed significantly between proficiency levels as well, $F(4,66) = 4.072$, $p = .005$, $\eta^2 = 0.198$. As Fig. 3 shows, the increases in MLC do not follow a linear pattern across proficiency levels. In the lowest two proficiency levels, the average length of the clauses produced by learners is around five and a half words. This value rises to almost six words in level 6 and slightly over six and a half in level 7. In level 8, MLC values drop again to the level of level 6. The pairwise comparisons show that MLC values are considerably higher in level 7 than in levels 4, 5 and the adjacent level 6. No sizeable difference was found between levels 7 and 8.

For the proportion of coordinated clauses we did not observe a significant overall difference between proficiency levels, $F(4,66) = 2.478$, $p = .052$, $\eta^2 = 0.130$. Slightly over one out of five clauses produced by the learners in level 4 was coordinated (see Fig. 4). This proportion rises to around 28% in level 6, but then drops, until it reaches its lowest point in level 8 (around 18%). Only two sizeable differences between proficiency levels were observed (4 *vs* 6 and 6 *vs* 8).

For the last syntactic complexity measure, mean length of noun phrase, the overall difference between proficiency levels was significant, $F(4,66) = 10.589$, $p < 0.001$, $\eta^2 = 0.391$. Learners in level 4 produce noun phrases with on average around 2.2 words (see Fig. 5). In levels 5 and 6, this value has risen to around 2.4, and in levels 7 and 8 it is close to three. Six significant differences be-

[8] As indicated by the large confidence interval for this proficiency level in Fig. 1.

[9] For reference, with 10 pairwise comparisons, applying Bonferroni's correction would amount to dividing the threshold for significance (α level) by 10 (e.g. $\alpha = 0.01$ becomes $\alpha = 0.001$).

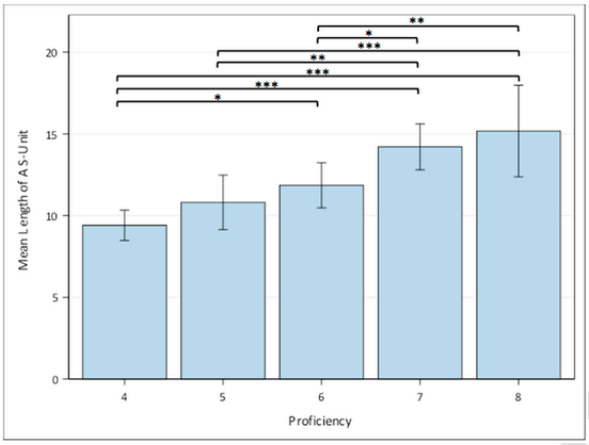

Fig. 1. Mean length of AS-unit scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: $p < .05$; **: $p < .01$; ***: $p < .001$).

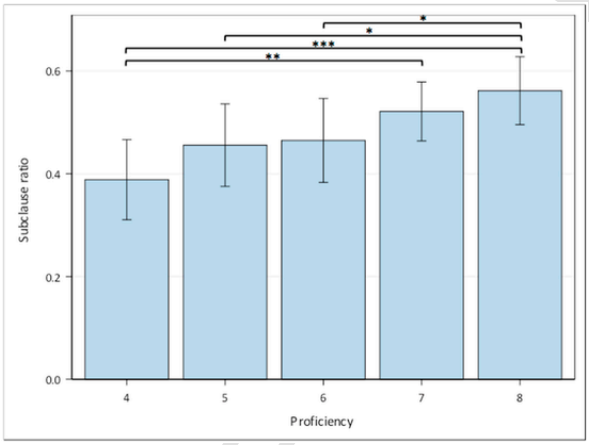

Fig. 2. Subclause ratio scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: $p < .05$; **: $p < .01$; ***: $p < .001$).

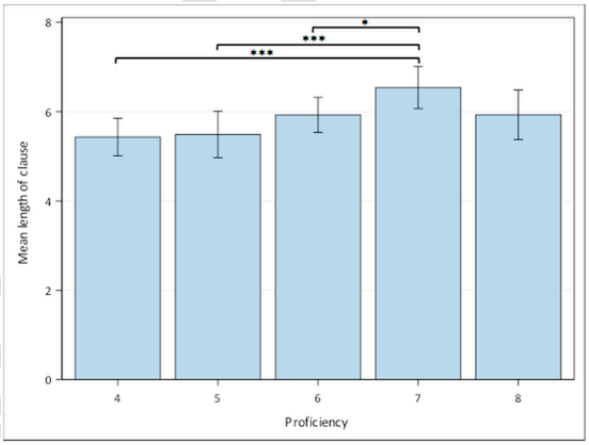

Fig. 3. Mean length of clause scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: $p < .05$; **: $p < .01$; ***: $p < .001$).

tween pairs of proficiency levels were found. The pairwise comparisons clearly show that MLNP scores can be divided into two distinct 'groups': levels 4, 5 and 6 on the one hand, and levels 7 and 8 on the other.

Turning to measures of lexical complexity, for Guiraud's index a significant overall effect of proficiency level was found, $F(4,66) = 15.162$, $p < 0.001$, $\eta^2 = 0.479$. As Fig. 6 shows, the learners in levels 4 and 5 obtain scores of around 5, and this value rises with approximately 0.5 points between the remaining consecutive levels to around 6.5 in level 8. In total, eight sizeable differ-

ences were found between proficiency levels. We found differences for two adjacent pairs of proficiency levels (5 *vs* 6 and 6 *vs* 7). Both levels 4 and 5 differ considerably from all other proficiency levels, and also the difference between levels 6 and 8 is significant.

The overall pattern for HD-D, the second lexical complexity measure, is fairly similar to the one we observed for Guiraud's index (see Fig. 7). Also here the overall effect of proficiency level is significant, $F(4,66) = 9.997$, $p < 0.001$, $\eta^2 = 0.377$, even though the effect size is considerably smaller. HD-D scores consistently rise from 0.65 to 0.72 from level 4 to level 8. In contrast to Guiraud's index, no considerable difference was found between levels 5 and 6, but all other pairs that yielded significant differences for Guiraud's index, also do so for HD-D.

For MTLD, the final lexical complexity measure, we also found a significant overall effect of proficiency level on the scores, $F(4,66) = 7.058$, $p < 0.001$, $\eta^2 = 0.300$. The lower effect size compared to Guiraud's index and HD-D is also reflected in the (slightly) fewer and less strongly significant pairwise comparisons. For MTLD, we did not find considerable differences between adjacent levels, in spite of consistently rising scores with rising proficiency level (increasing from around 25 in level 4 to around 37 in level 8; see Fig. 8).

Finally, we look at the results for MCI, our measure of morphological complexity. The overall difference between proficiency levels is significant, $F(4,66) = 6.804$, $p < 0.001$, $\eta^2 = 0.292$. It is clear from Fig. 9 that the observed difference mainly lies between the lowest proficiency level and all other levels combined. The only considerable pairwise differences are observed between level 4 and all other levels. This means that also the difference between the adjacent levels 4 and 5 is significant.

5.2. Which complexity measures correlate best with L2 proficiency?

In order to evaluate which complexity measures show the strongest effect of L2 proficiency level, we calculated Spearman correlation coefficients between the scores on each measure and the proficiency level of the learners. Table 3 shows that the correlation is significant (and positive) for eight out of nine measures, with the strongest correlations found for Guiraud's index ($\rho = 0.701$), followed by two syntactic measures (MLNP and MLAS, with $\rho = 0.616$ and $\rho = 0.615$) and another lexical measure (HD-D; $\rho = 0.615$). Only for the coordination ratio the correlation coefficient was not significant. This is not surprising, since the scores on this measure follow an inverted-V-shaped rather than a linear pattern (see Fig. 4).

Appendix C shows the Pearson correlations between the different complexity measures. Many of these correlations are positive and significant, which, at least in part, can be attributed to their shared correlation with proficiency. The highest correlations are found between the lexical complexity measures HD-D and G ($r = 0.886$) and HD-D and MTLD ($r = 0.839$), which all target lexical diversity. Also the correlation between the syntactic measures MLAS and SCR is high ($r = 0.811$), which is not surprising considering the structural interconnectedness of these two measures (i.e. AS-units can be made longer by including more subclauses). We also found two significantly negative correlations, between CCR on the one hand and SCR ($r = -0.452$) and MLAS ($r = -0.312$) on the other, reflecting that speakers have to choose between coordination and subordination as clause-linking mechanism.

Considering its high correlation with both G and MTLD, we did not include HD-D in the ordinal logistic regression model with proficiency level as dependent variable. Also MLAS was left out due to high correlations with SCR and, to a somewhat lesser extent, MLNP. Finally, we also did not include CCR, since this measure was found to have a clearly non-linear relationship with proficiency. The score test does not indicate that the proportional odds assumption of our cumulative logit model is violated, $\chi^2 = 27.86$, $df = 18$, $p = 0.064$. Nagelkerke's R^2 of 0.676 indicates that around two thirds of the variance in proficiency level is explained by the model. Table 4 contains the parameter estimates for the ordinal logistic regression model.[10]

The model contains three significant predictors, G, SCR and MCI, each targeting complexity in a different linguistic domain (i.e. lexical, syntactic and morphological). Of these three, G is the strongest predictor ($\beta = -1.04$), followed by SCR ($\beta = -0.63$) and MCI ($\beta = -0.38$). The contributions of the three remaining predictors (MLC, MLNP and MTLD) are not significant in the model.

6. Discussion

6.1. Can complexity measures discriminate between proficiency levels?

Generally speaking, the results of this study confirm that L2 speech complexity measures can discriminate between learners at different proficiency levels, even though we also observed considerable differences across measures and across proficiency levels. Looking more closely at the results per complexity dimension, our finding that morphological complexity could only differentiate between the lowest level and all higher levels is in line with De Clercq and Housen (2019). A potential explanation for this is the limited range of inflectional morphology affecting the verbal system in present-day English, resulting in a limited number of possible forms (Brezina & Pallotti, 2019). Nevertheless, morphological complexity, and the morphological complexity index in particular, appears to be an appropriate indicator of English L2 proficiency at early stages of development.

Regarding lexical complexity, our study confirms that lexical diversity measures (Guiraud's index and HD-D in particular) are able to discriminate between (adjacent) proficiency levels (De Clercq, 2015; Lu, 2012). The three lexical diversity measures included in this study showed a gradual increase between levels 4 to 8 of the IELTS speaking test, which corroborates Read and Nation's (2006)

[10] In accordance with the default model specification in SAS, negative parameter estimates for a predictor indicate that higher scores for the corresponding complexity measure are associated with higher proficiency levels.

Table 3
Correlations between complexity measures and proficiency level (*: p < .05; **: p < .01; ***: p < .001).

	Spearman correlation coefficient (ρ)
Syntactic complexity	
Mean length of AS-unit	0.616***
Subclause ratio	0.419***
Mean length of clause	0.301*
Coordination ratio	−0.057
Mean length of NP	0.615***
Lexical complexity	
Guiraud's index	0.701***
HD-D	0.615***
MTLD	0.535***
Morphological complexity	
MCI-verbs	0.526***

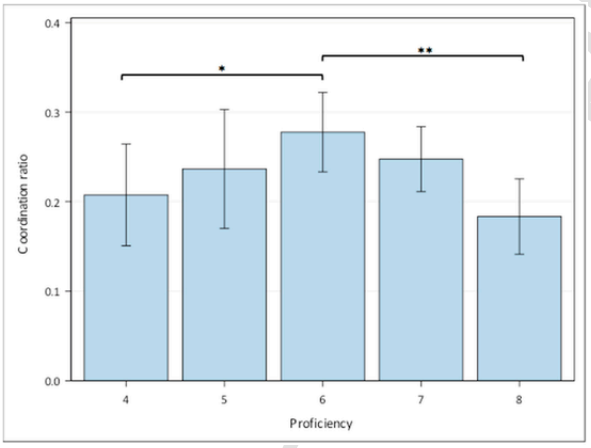

Fig. 4. Coordination ratio scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: p < .05; **: p < .01; ***: p < .001).

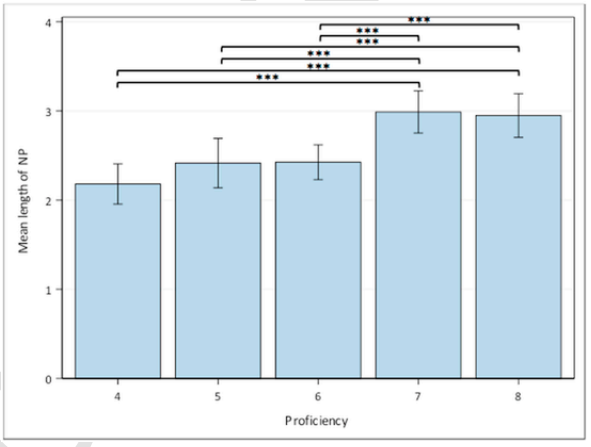

Fig. 5. Mean length of NP scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: p < .05; **: p < .01; ***: p < .001).

findings. Moreover, we found that all diversity measures gave rise to significant differences between proficiency levels, but only Guiraud's index and HD-D could discriminate between some adjacent levels: levels 5 and 6, and 6 and 7 for Guiraud's index, and only levels 6 and 7 for HD-D. As in De Clercq (2015), there were no significant differences in lexical complexity between the two highest levels (7 and 8, in our study). This, however, contrasts with Kang (2013), who concluded that differences in lexical complexity mainly manifested themselves between CEFR levels B1 and B2, and C1 and C2, which can be said to correspond to IELTS bands 7 and 8. It should be noted here that Kang (2013) also included measures of lexical density and sophistication.

Four of our five syntactic complexity measures (i.e. mean length of AS-unit, clause and noun phrase, and the subclause ratio) showed an overall significant effect for proficiency level, whereas the coordination ratio did not. The fact that the measure of supra-

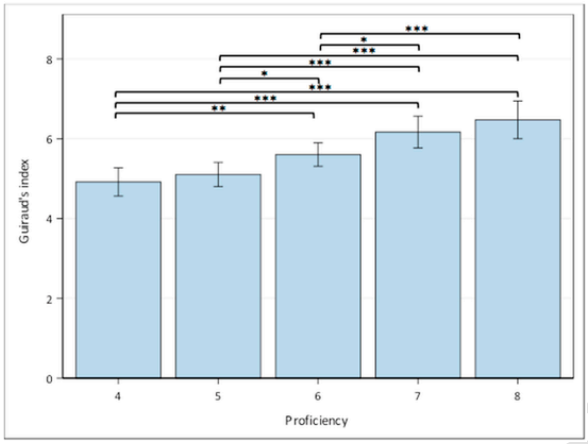

Fig. 6. Guiraud scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: p < .05; **: p < .01; ***: p < .001).

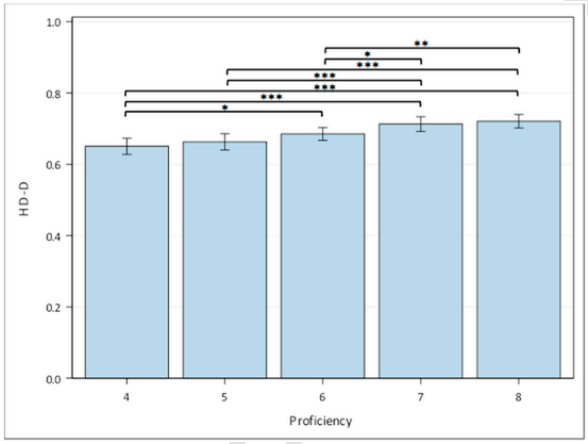

Fig. 7. HD-D scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: p < .05; **: p < .01; ***: p < .001).

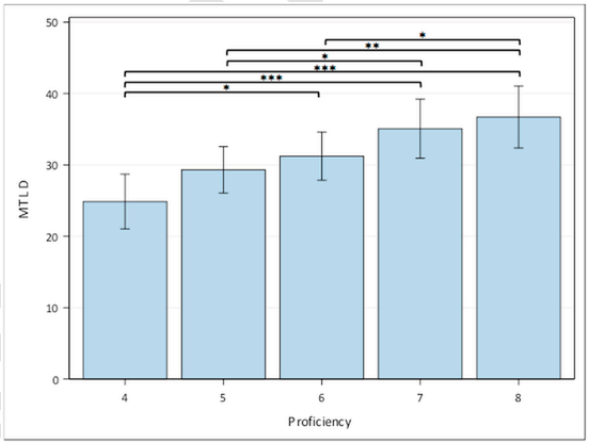

Fig. 8. MTLD scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: p < .05; **: p < .01; ***: p < .001).

clausal complexity (i.e. mean length of AS-unit) showed the highest number of significant differences between proficiency levels is in line with previous studies (De Clercq & Housen, 2017; Gan, 2012; Iwashita et al., 2008). The significant differences between proficiency levels we found for mean length of noun phrase, and to a lesser extent mean length of clause, are more unexpected (De Clercq & Housen, 2017), even though it should be added that not many previous studies included measures of clausal and phrasal complexity. In the present study, mean length of AS-unit, clause and noun phrase not only resulted in an overall significant difference between proficiency levels, but they could also distinguish between two adjacent levels, levels 6 and 7. In De Clercq and Housen'

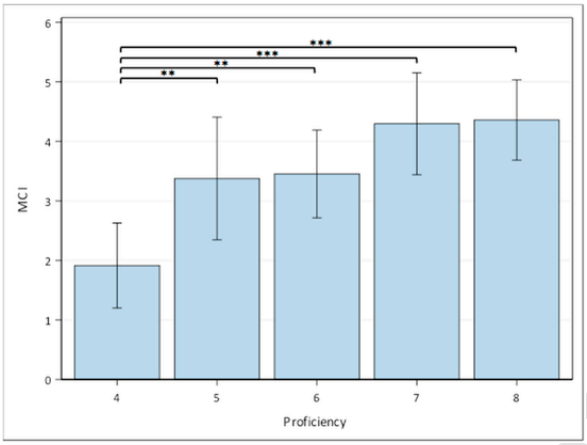

Fig. 9. MCI-verbs scores per proficiency level, with significant ANOVA post-hoc pairwise differences (*: p < .05; **: p < .01; ***: p < .001).

Table 4
Parameter estimates for ordinal logistic regression.

Parameter	Estimate	Standard Error	Standardised estimate (β)	Wald χ^2	p-value
Intercepts					
4	15.11	2.96		26.09	<0.0001***
5	16.94	3.06		30.54	<0.0001***
6	18.81	3.21		34.28	<0.0001***
7	20.82	3.41		37.28	<0.0001***
SCR	−8.33	2.32	−0.63	12.86	0.0003***
MLC	0.32	0.34	0.16	0.89	0.3452
MLNP	−1.05	0.60	−0.30	3.01	0.0826
G	−2.19	0.53	−1.04	16.81	<0.0001***
MTLD	0.03	0.04	0.11	0.33	0.5667
MCI	−0.42	0.20	−0.38	4.32	0.0376*

*: p < .05; **: p < .01; ***: p < .001. Nagelkerke's R^2 = 0.676; -2LL = 154.14; Somer's D = 0.748; ROC = 0.874.

s (2017) data, mean length of AS-unit could distinguish between levels 2 and 3 of the four proficiency levels used in the study, but not between the third and the most advanced level, which can be considered to be similar in our data, where none of the measures could distinguish between the two highest levels. The finding that the subclause ratio gave rise to an overall significant difference between levels confirms Seedhouse et al.'s (2014) study, which also used IELTS speaking test data, and is also in line with other previous studies (De Clercq & Housen, 2017; Kang, 2013). While Seedhouse et al. (2014) did not investigate between which levels the differences could be found, our pairwise comparisons show that even though subordination increases linearly from levels 4 to 8, it cannot discriminate between adjacent levels.

6.2. Which complexity measures correlate best with proficiency?

The results of the ordinal logistic regression are novel in comparison to the previous studies on the relationship between L2 speech complexity and L2 proficiency which we described. They show that one lexical (Guiraud's index), one syntactic (subclause ratio) and one morphological complexity measure (morphological complexity index - verbs) emerged as significant predictors of L2 proficiency, together explaining an estimated two thirds of the variance in proficiency level. This analysis indicates that it is worthwhile to combine different complexity dimensions and measures tapping into various aspects of complexity when the aim is to assess L2 proficiency. Generally speaking, Guiraud's index of lexical diversity emerged from our study as the best overall predictor of L2 proficiency. Also in Verspoor et al.'s (2012) study of L2 writing, Guiraud's index was shown to be one of the measures that could discriminate best between learners with different proficiency levels, and this across a wide range of levels. Our detailed between-level analyses for the different measures also show, however, that the relationship between complexity measures and L2 proficiency is not always linear, and that different measures may be more sensitive to changes in proficiency at different proficiency levels.

6.3. Limitations

This study has a number of limitations. First, the IELTS data we used are not longitudinal, which means we could only characterise L2 developmental patterns in broad strokes (at the group level) and not make any claims about individual L2 develop-

ment. This implies we had to disregard individual variation to a large extent, as well as individual developmental patterns, which have been shown to vary greatly across learners (Bulté & Housen, 2018; Vyatkina et al., 2015). Second, no corrections for multiple testing (e.g. Bonferroni) were applied to the pairwise comparisons accompanying the ANOVAs, even though this is often recommended (Field, 2009). Future studies could aim for larger sample sizes per proficiency level in order to increase the statistical power of the tests. Increasing the sample size would also be beneficial for the generalisability of the results. Third, even though the score test indicated that the proportional odds assumption underlying our cumulative logit model was not violated, such a model assumes equal effects of dependent variables across different levels of the independent variable. The descriptive statistics indicate that such a model cannot fully do justice to the diverse patterns observed for the different complexity measures. Finally, proficiency level is operationalised in this study as IELTS level, which is determined on the basis of subjective ratings of speaking tasks, including the monologues that constitute our dataset. Moreover, the rubrics used for the subjective assessment of proficiency level include references to complexity-related factors. It cannot be ruled out that at least part of the association between proficiency level and complexity scores found in this study is caused by this connection.

7. Conclusions

The present study on the relationship between nine complexity measures and five different levels of oral proficiency, as measured by the IELTS speaking test, confirms previous studies which have found that learners at higher levels of proficiency tend to produce more complex language. Even though we found higher complexity scores in higher proficiency levels for measures of lexical, syntactic and morphological complexity, the observed patterns differ substantially across measures. If we only consider differences between adjacent proficiency levels, we observed a significant increase in morphological richness (as measured by the morphological complexity index) between levels 4 and 5, in lexical diversity (Guiraud's index) between levels 5 and 6, and in overall syntactic (mean length of AS-unit), clausal (mean length of clause) and phrasal complexity (mean length of noun phrase) as well as lexical diversity (Guiraud's index and HD-D) between levels 6 and 7. We did not observe significant differences in complexity between the highest two proficiency levels in our dataset (i.e. 7 and 8). In addition, we found that the Guiraud index, the subclause ratio and the morphological complexity index applied to verbs were significant predictors for proficiency level in our ordinal logistic regression model, explaining around two thirds of the variance in proficiency level.

With this study, we did not intend to make claims about the inherent value or validity of complexity measures as measures of complexity. Rather, our aim was to find out which complexity measures show significant differences between learners with different proficiency levels or, in other words, which measures are potentially good indicators of L2 proficiency (and, potentially, L2 development). In this sense, our results are informative for future studies using complexity measures as (objective) indicators of L2 learner proficiency. They also have implications for language testing, even though the labour-intensiveness of transcribing and coding oral data are an obstacle to automated analyses, since current automated complexity tools are geared towards the analysis of written language, or thoroughly cleaned up and carefully transcribed versions of oral data. Nevertheless, objective complexity analyses can be a useful addition to subjective proficiency ratings, also for language testing purposes.

Funding

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

CRediT authorship contribution statement

Bram Bulté: Conceptualization, Methodology, Software, Formal analysis. **Hanne Roothooft:** Conceptualization, Formal analysis, Investigation, Resources, Data curation.

Acknowledgements

This paper reports on research using data provided by Cambridge English Language Assessment.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.system.2020.102246.

Appendix A. Descriptive statistics

Table A.1
Descriptive statistics (means with standard deviations between brackets).

	Proficiency level					
	4 (n = 14)	5 (n = 14)	6 (n = 15)	7 (n = 14)	8 (n = 14)	Total (n = 71)
Syntactic complexity						

Mean length of AS-unit	9.40	10.81	11.86	14.20	15.17	12.28
	(1.60)	(2.89)	(2.49)	(2.43)	(4.85)	(3.64)
Subclause ratio	38.86	45.57	46.50	52.13	56.17	47.83
	(13.50)	(13.91)	(14.74)	(9.94)	(11.45)	(13.82)
Mean length of clause	5.43	5.49	5.93	6.54	5.93	5.86
	(0.73)	(0.90)	(0.71)	(0.82)	(0.96)	(0.90)
Coordination ratio	20.75	23.66	27.76	24.75	18.34	23.12
	(9.83)	(11.50)	(7.99)	(6.29)	(7.30)	(9.13)
Mean length of NP	2.18	2.42	2.43	2.99	2.95	2.59
	(0.39)	(0.35)	(0.35)	(0.41)	(0.42)	(0.51)
Lexical complexity						
Guiraud's index	4.92	5.11	5.60	6.16	6.47	5.65
	(0.61)	(0.52)	(0.53)	(0.69)	(0.81)	(0.86)
HD-D	0.65	0.66	0.69	0.71	0.72	0.69
	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)
MTLD	24.87	29.32	31.22	35.07	36.71	31.44
	(6.64)	(5.63)	(6.07)	(7.16)	(7.53)	(7.70)
Morphological complexity						
MCI	1.91	3.38	3.45	4.30	4.36	3.48
	(1.24)	(1.79)	(1.33)	(1.48)	(1.17)	(1.64)

Appendix B. ANOVAs

Table B.1
Results ANOVAs.

	F(4,66)	p	Effect size (eta ^{2})
Syntactic complexity			
Mean length of AS-unit	8.574	0.000***	0.342
Subclause ratio	3.714	0.009**	0.184
Mean length of clause	4.072	0.005**	0.198
Coordination ratio	2.478	0.052	0.130
Mean length of NP	10.589	0.000***	0.391
Lexical complexity			
Guiraud's index	15.162	0.000***	0.479
HD-D	9.997	0.000***	0.377
MTLD	7.058	0.000***	0.300
Morphological complexity			
MCI	6.804	0.000***	0.292

Appendix C. Correlations between complexity measures

Table C.1
Pearson correlations between complexity measures (*: p < .05; **: p < .01; ***: p < .001)

	SCR	MLC	CCR	MLNP	MCI	G	HD-D	MTLD
MLAS	0.811***	0.340**	−0.312**	0.605***	0.232	0.384***	0.377**	0.474***
SCR		−0.100	−0.452***	0.373**	−0.037	0.092	0.094	0.258*
MLC			0.151	0.436***	0.488***	0.540***	0.488***	0.438***
CCR				−0.078	0.188	0.172	0.143	0.016
MLNP					0.376**	0.486***	0.419***	0.456***
MCI						0.646***	0.587***	0.492***
G							0.886***	0.682***
HD-D								0.839***

References

Abbas Khushik, G, & Huhta, A (2019). Investigating syntactic complexity in EFL learners' writing across common European framework of reference levels A1, A2, and B1. Applied Linguistics. doi:10.1093/applin/amy064. amy064.
Agresti, A (2010). Analysis of ordinal categorical data (2nd ed.). New York: Wiley.
Biber, D (1988). Variation across speech and writing. Cambridge: Cambridge University Press.
Biber, D, Gray, B, & Poonpon, K (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? Tesol Quarterly, 45(1), 5–35.
Biber, D, Gray, B, & Staples, S (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. Applied Linguistics, 37(5), 639–668.
Brezina, V, & Pallotti, G (2019). Morphological complexity in written L2 texts. Second Language Research, 35(1), 99–119.

Bulté, B, & Housen, A (2012). Defining and operationalising L2 complexity. In Housen, A, Kuiken, F, & Vedder, I (Eds.), Dimensions of L2 performance and proficiency. Investigating complexity, accuracy and fluency in SLA (pp. 21–46). Amsterdam: John Benjamins.
Bulté, B, & Housen, A (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. Journal of Second Language Writing, 23(4), 21–45. doi:10.1016/j.jslw.2014.09.005.
Bulté, B, & Housen, A (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. International Journal of Applied Linguistics, 28(1), 1–18. doi:10.1111/ijal.12196.
Bulté, B, Housen, A, Pierrard, M, & Van Daele, S (2008). Investigating lexical proficiency development over time – the case of Dutch-speaking learners of French in Brussels. Journal of French Language Studies, 18, 277–298. doi:10.1017/S0959269508003451.
Connor-Linton, J, & Polio, C (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. Journal of Second Language Writing, 23, 1–9.
Crosthwaite, P (2016). A longitudinal multidimensional analysis of EAP writing: Determining EAP course effectiveness. Journal of English for Academic Purposes, 22, 166–178.
De Clercq, B (2015). The development of lexical complexity in second language acquisition. A cross-linguistic study of L2 French and English. EUROSLA Yearbook, 15, 69–94.
De Clercq, B, & Housen, A (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. The Modern Language Journal, 101(2), 315–334.
De Clercq, B, & Housen, A (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. Second Language Research, 35(1), 71–97.
Derr, R (2013). Ordinal response modeling with the LOGISTIC procedure. Proceedings of the SAS global forum 2013 conference. Cary, NC: SAS Institute Inc.
Ellis, R, & Barkhuizen, G P (2005). Analysing learner language. Oxford: Oxford University Press.
Field, A (2009). Discovering statistics using SPSS (3rd ed.). London: Sage.
Foster, P, Tonkyn, A, & Wigglesworth, G (2000). Measuring spoken language: A unit for all reasons. Applied Linguistics, 21(3), 354–375.
Friginal, E, & Weigle, S (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. Journal of Second Language Writing, 26, 80–95.
Gan, Z (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. Language Assessment Quarterly, 9(2), 133–151.
Guiraud, P (1959). Problèmes et méthodes de la statistique linguistique. Dordrecht: Reidel.
Gyllstad, H, Granfeldt, J, Bernardini, P, & Källkvist, M (2014). Linguistic correlates to communicative proficiency levels of the CEFR. The case of syntactic complexity in written L2 English, L3 French and L4 Italian. EUROSLA Yearbook, 14, 1–30.
Hawkins, J A, & Filipović, L (2012). Criterial features in L2 English: Specifying the reference levels of the Common European framework: 1. Cambridge: Cambridge University Press.
Horst, M, & Collins, L (2006). From faible to strong: How does their vocabulary grow? Canadian Modern Language Review, 63(1), 83–106.
Housen, A, Kuiken, F, & Vedder, I (Eds.) (2012). Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA: 32. Amsterdam: John Benjamins Publishing.
Hulstijn, J H (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. Language Assessment Quarterly, 8(3), 229–249.
Iwashita, N (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. Language Assessment Quarterly, 3(2), 151–169.
Iwashita, N, Brown, A, McNamara, T, & O'Hagan, S (2008). Assessed levels of second language speaking proficiency: How distinct? Applied Linguistics, 29(1), 24–49.
Jarvis, S (2013). Capturing the diversity in lexical diversity. Language Learning, 63(Suppl. 1), 87–106.
Kang, O (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. Cambridge English: Research Notes, 52, 40–48.
Leclercq, P, Edmonds, A, & Hilton, H (Eds.) (2014). Measuring L2 proficiency: Perspectives from SLA. (78). Clevedon: Multilingual Matters.
Lu, X (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. Tesol Quarterly, 45(1), 36–62.
Lu, X (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. The Modern Language Journal, 96(2), 190–208.
Malvern, D, Chipere, N, Richards, B, & Durán, P (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills. Palgrave Macmillan.
McCarthy, P, & Jarvis, S (2007). Vocd: A theoretical and empirical evaluation. Language Testing, 24(4), 459–488.
McCarthy, P, & Jarvis, S (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381–392.
Michel, M (2017). Complexity, accuracy, and fluency in L2 production. In Loewen, S, & Sato, M (Eds.), The routledge handbook of instructed second language acquisition (pp. 50–68). London: Routledge.
Neary-Sundquist, C A (2017). Syntactic complexity at multiple proficiency levels of L2 German speech. International Journal of Applied Linguistics, 27(1), 242–262.
Ortega, L (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics, 24(4), 492–518.
Ortega, L (2012). Interlanguage complexity: A construct in search of theoretical renewal. In Kortmann, B, & Szmrecsanyi, B (Eds.), Linguistic complexity: Second language acquisition, indigenization, contact (pp. 127–155). Berlin: de Gruyter.
Pallotti, G (2009). CAF: Defining, refining and differentiating constructs. Applied Linguistics, 30(4), 590–601.
Pallotti, G (2015). A simple view of linguistic complexity. Second Language Research, 31(1), 117–134.
Read, J, & Nation, P (2006). An investigation of the lexical dimension of the IELTS speaking test. International English Language Testing System (IELTS) Research Reports, 6, 1. 2006.
Rescher, N (1998). Complexity: A philosophical overview. London: Transaction Publishers.
Seedhouse, P, Harris, A, Naeb, R, & Üstünel, E (2014). The relationship between speaking features and band descriptors: A mixed method study. IELTS Research Reports Online Series, 2, 1–30.
Spoelman, M, & Verspoor, M (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. Applied Linguistics, 31, 532–553.
Thomas, M (1994). Assessment of L2 proficiency in second language acquisition research. Language Learning, 44, 307–336.
Vasylets, O, Gilabert, R, & Manchón, R M (2019). Differential contribution of oral and written modes to lexical, syntactic and propositional complexity in L2 performance in instructed contexts. Instructed Second Language Acquisition, 3(2), 206–227.
Verspoor, M, Schmid, M, & Xu, X (2012). A dynamic usage based perspective on L2 writing. Journal of Second Language Writing, 21, 239–263.
Vyatkina, N, Hirschmann, H, & Golcher, F (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. Journal of Second Language Writing, 29, 28–50.
Wolfe-Quintero, K, Inagaki, S, & Kim, H-Y (1998). Second language development in writing: Measures of fluency, accuracy, and complexity. Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center.
Yang, W, Lu, X, & Weigle, S (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. Journal of Second Language Writing, 28, 53–67.
Yoon, H (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. System, 66, 130–141.